Univerzita Karlova

Filozofická fakulta
Fonetický ústav

DIPLOMOVÁ PRÁCE

Bc. Dita Lazárková

Využití dlouhodobé formantové distribuce pro rozpoznatelnost mluvčího v různých akustických podmínkách

Using long-term formant distributions for speaker identification in various acoustic conditions

2015

vedoucí práce: Mgr. Radek Skarnitzl, Ph.D.

Ráda bych zde poděkovala svému vedoucímu, Mgr. Radku Skarnitzlovi, PhD., za veškerou pomoc, konzultace, rady, a především vstřícnost při přípravě této diplomové práce. Další díky patří Ing. Ondřeji Břížkovi a Zbyňku Lazárkovi za pomoc s řešením různých technických a statistických problémů. Samozřejmě nesmím zapomenout na všechny dobrovolníky, kteří mi věnovali svůj čas a uvolili se k pořízení nahrávek.

Děkuji.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze 12. ledna 2015

.....

Bc. Dita Lazárková

Abstrakt

Analýza dlouhodobé formantové distribuce (LTF) je poměrně mladou, ale slibnou disciplínou identifikace mluvčích. Jde o metodu mapující dlouhodobé chování formantů v řeči jednotlivých mluvčích. Častými problémy, s nimiž se v praxi setkáváme, je zhoršená akustická kvalita a příliš krátké trvání zkoumaných nahrávek. Tato práce má za cíl představit historický vývoj forenzní fonetiky a současné používané metody. V praktické části se zabýváme využitelností LTF metody ve forenzní praxi, zvláště pak u nahrávek obsahujících šum v pozadí. Ukázalo se, že šum extrahované LTF hodnoty znatelně ovlivňuje, bohužel nejde o žádné systematické změny. Proto jsme navrhli několik metod, jak šum v nahrávce kompenzovat, aby bylo možné navzájem srovnávat i čisté a zašuměné nahrávky. Zkoumali jsme též minimální trvání nahrávky, jež je nutné pro statistickou spolehlivost výsledných hodnot. Tato hranice není přesná a je pro jednotlivé mluvčí značně individuální, ale je patrné, že nahrávky (vokalické proudy) kratší než 15 s mají již často sníženou vypovídající hodnotu, nelze je tedy pro analýzu doporučit.

Klíčová slova: LTF, dlouhodobá formantová distribuce, identifikace mluvčích, forenzní fonetika, akustická kvalita nahrávky, šum

Abstract

The analysis of long-time formant distribution is relatively young but promising discipline of speaker identification. It is a method of mapping the long-term behavior of formants in speech of individual speakers. Frequently encountered problems in practice are bad acoustic quality and very short duration of analyzed recordings. This work aims to present the historical development of forensic phonetics and currently used methods. In the practical part, it deals with the usability of LTF method in forensic practice, especially in recordings containing background noise. It was shown that the noise appreciably affects extracted LTF values and unfortunately the change is not systematic. Therefore, we proposed several methods to compensate the noise in recordings, in order to be able to compare recordings with and without noise. We also investigated the minimum duration of recording, which is necessary for statistical reliability of the resulting values. This boundary is not exact and for each speaker, it is substantially individual. But it is apparent that recordings (vocalic streams) shorter than 15 s often provide incomplete information, wherefore they cannot be recommended for analysis.

Keywords: LTF, long-time formant distribution, speaker identification, forensic phonetics, acoustic quality of recording, noise

Obsah

Úvod

Identifikace mluvčího tvoří základní a nejdůležitější součást forenzní fonetiky. S vynálezem telefonu, nahrávacích zařízení, a především s rozvojem výpočetní techniky v posledních desetiletích se kromě tradiční laické poslechové metody začíná výrazně uplatňovat i metoda akustické analýzy, pod niž analýza dlouhodobých formantových distribucí (LTF) spadá. Akustická analýza umožňuje výzkum mnohých aspektů a vlastností lidského hlasu, jež nejsme schopni sluchem rozeznat a zhodnotit, ale které se významně podílejí na výsledné podobě a dojmu hlasu. Samotná LTF metoda je mimo jiné určena k rychlé předběžné kategorizaci mluvčích, abychom následné časově náročné analýzy a srovnávání prováděli již na užším výběru možných shodných mluvčích. Není proto vhodná k samostatnému využití, ale v kombinaci s dalšími technikami akustické i poslechové analýzy může být silným identifikačním nástrojem.

V první části práce představíme teoretické pozadí této studie – po seznámení s historickým vývojem forenzní fonetiky v 1. kapitole následuje ve 2. kapitole přehled oblastí působnosti forenzní fonetiky. 3. kapitolu tvoří úvod do procesu identifikace mluvčího a příklady možných analytických metod včetně představení samotné metody LTF a detailnějšího přiblížení cílů této studie. Druhá část práce je praktická a popisuje použitou metodu (4. kapitola) a závěry studie v souvislostech (5. a 6. kapitola).

Jelikož je ve forenzní praxi častým problémem špatná akustická kvalita nahrávek, zaměřujeme se v této práci na výzkum toho, zda je LTF metoda ve zhoršených podmínkách (za přítomnosti šumu v pozadí) dostatečně robustní a zda přítomné šumy nějakým způsobem ovlivňují výsledky metody. Pokusíme se navrhnout možné kompenzační techniky, které by umožnily srovnávání nahrávek pořízených za různých akustických podmínek, a zjistit minimální trvání nahrávky potřebné pro spolehlivou extrakci LTF distribuce.

I Teoretická část

1 Historický vývoj forenzní fonetiky

Ačkoliv obor forenzní fonetiky zažívá největší rozvoj v posledních desetiletích, laická identifikace pachatele podle jeho hlasu se pravděpodobně v soudnictví využívá po staletí, ne-li tisíciletí. Jeden z nejstarších a nejznámějších doložených případů (Eriksson, 2005, s. 2) pochází z roku 1660, kdy byl veden soudní proces s Williamem Huletem jako domnělým popravčím anglického krále Karla I. Svědek Richard Gittens přísahal, že Huleta poznal pod popravčí kápí podle hlasu, na základě čehož měl být Hulet popraven. Ještě před tím však vyšlo najevo, že popravu provedl jiný kat, který se k ní i přiznal a Hulet byl osvobozen. Bohužel se tak zároveň jedná i o jeden z prvních známých případů špatné identifikace mluvčího.

Že identifikace pachatele pouze na základě sluchového svědectví může být – zvláště s prodlužující se dobou mezi samotným trestným činem a soudním jednáním – dost zpochybnitelná, se ukázalo i v případu únosu z roku 1932. Tehdy byl z únosu syna Charlese Lindbergha obviněn německý přistěhovalec Bruno Hauptmann, jehož v soudním řízení po více než dvou letech po únosu Lindbergh identifikoval jen kvůli jeho silnému německému přízvuku (Moos, 2008, s. 10). Na základě tohoto případu vypracovala Francis McGehee první dvě empirické studie (McGehee, 1937; 1944) zabývající se zpětnou rozpoznatelností hlasu s časovým odstupem. Subjekty prováděly identifikaci po 1, 2 a 3 dnech, po 1, 2 a 3 týdnech a po 1, 3 a 5 měsících od prvního poslechu. Po jednom dnu či týdnu se úspěšnost rozpoznání mluvčího pohybovala kolem 80 %. Po dvou týdnech ale klesla na 69 %, po měsíci na 57 %, po třech měsících na 35 % a po pěti měsících na 13 %, což bylo dokonce pod hranicí náhody. Ve druhém experimentu posuzovali posluchači místo živých hlasů nahrávky, ale výsledky zůstaly obdobné.

Nové technické vynálezy jako telefon či nahrávací zařízení otevřely ve forenzní praxi nové možnosti. Významným pracovištěm byly od 30. let 20. století Bell

Telephone Laboratories, kde vznikl první spektrogram (grafické zobrazení zvuku na základě frekvence, intenzity a času). Zhruba až do konce 2. světové války zůstal spektrogram ovšem nevyužitý, neboť se jednalo o válečný projekt Spojených států. Přitom deklarovaným cílem výzkumu bylo usnadnění výuky řeči hluchým mluvčím či studentům cizího jazyka, což je dle Erikssona (2005, s. 3) přinejmenším zvláštní. Meuwly (2003) se domnívá, že skutečným cílem bylo vytvořit metodu identifikace mluvčích. V první publikované práci použili Bellovi výzkumníci Grey a Kopp (1944) dokonce termín „voiceprint", tedy „hlasový otisk", jako metaforu vycházející z otisků prstů, které jsou pro každého člověka unikátní. Z románu Alexandra Solženicyna *V kruhu prvním* vyplývá, že obdobný výzkum možnosti identifikace mluvčího prováděli i uvěznění vědci v Sovětském svazu (Eriksson, 2005, s. 3). Z použité terminologie lze přitom usuzovat, že s probíhajícím vývojem na Západě byli obeznámeni.

Po konci války se Bellovy laboratoře vrátily opět k tradičním fonetickým tématům, k výuce řeči a terapii. Až do roku 1962 nejsou žádné zmínky o vývoji v identifikaci mluvčích. V roce 1962 publikoval Lawrence Kersta v časopise Nature práci pojmenovanou *Voiceprint identification*. Kersta byl zaměstnancem Bellových laboratoří, je tedy pravděpodobné, že tam s výzkumem využití spektrogramu jako nástroje pro identifikaci mluvčích přišel do styku. V této práci představil revoluční názor, že je možné pouze na základě vizuálního posouzení spektrogramu identifikovat mluvčího s úspěšností minimálně 99 %. Vycházel z předpokladu, že žádní dva lidé nemají zcela stejnou anatomii vokálního traktu nebo naučené artikulační strategie, a nemůžou mít tedy ani stejné spektrogramy, které z těchto fyzických faktorů vychází. V roce 1966 odešel z laboratoře a začal sám cvičit odborníky na identifikaci pomocí spektrogramu. Zpočátku slavil s metodou velký úspěch, ale po několika letech se začaly objevovat mnohé kritické reakce. Kersta se totiž domníval, že je skutečně možné přiřadit ke každému spektrogramu právě jednoho mluvčího, tedy že spektrogram funguje na stejné bázi jako např. otisk prstu či analýza DNA (termín „voiceprint" tu už není metaforou, ale je brán doslova). Proti tomu se ozývali různí odborníci s odůvodněním, že nejen že více mluvčích může mít stejný spektrogram, ale dokonce i jeden mluvčí může za různých podmínek produkovat odlišné spektrogramy týchž slov. Zastáncem Kerstovy metody byl např. O. Tosi, podle jehož studie (Tosi et al., 1972) se chybovost identifikace mluvčích metodou voiceprintu pohybovala v rozmezí 5–15 % v závislosti na podmínkách. Oproti tomu ve studii, kterou provedli Young a Campbell (1967), dosáhli hodnotitelé na izolovaných slovech jen 78 % úspěšnosti,

na slovech z různých kontextů pak dokonce jen 38 %. Obdobné výsledky přinesla i studie Stevense et al. (1968). Hodnotitelé měli za úkol identifikovat mluvčí jednak poslechem ze sluchátek, jednak vizuálně ze spektrogramů. Chybovost při poslechové identifikaci dosáhla 6 %, při vizuálním hodnocení ovšem až 21 %. Téma identifikace mluvčího metodou voiceprintu rozdělovalo odborníky až do konce 80. let, kdy soudy přestaly tuto metodu považovat za průkaznou. Ještě dnes prý však tento postup používají někteří soukromí detektivové.

Kontroverze metody voiceprintu bohužel negativně ovlivnila náhled i na další fonetické metody a jejich využitelnost při forenzní identifikaci mluvčích a mnoho odborníků raději od tématu identifikace ve své práci upustilo. Přitom některé součásti metody byly určitě krokem vpřed – pozitivní např. bylo, že extrakce a porovnávání parametrů byly založené na systematické a kvantitativní bázi (viz Nolan, 1999). Následky kauzy se promítly i do samotné forenzní praxe – řada soudů v USA zvažovala, zda je vůbec svědectví na základě srovnávání hlasu mluvčích možné přijmout (Tiersma & Solan, 2002).

Koncem 60. let 20. století se začaly objevovat studie zkoumající možnost identifikace mluvčího na základě segmentů řeči (hlásek). Např. Glenn a Kleiner (1968) a Wolf (1972) představili slibné výsledky ohledně individuální výslovnosti nazál. Bohužel právě nazály jsou hodně náchylné na zdravotní stav mluvčího (např. nachlazení), a nejsou tedy z dlouhodobého hlediska invariantní.

Asi v polovině 70. let 20. století se začíná objevovat myšlenka automatické identifikace mluvčích. Jelikož je řeč závislá na velkém množství faktorů (z nichž některé pravděpodobně ještě ani neznáme) a neexistuje žádný rys, podle nějž samotného by bylo možné mluvčí rozpoznávat, zcela automatická identifikace mluvčích podle hlasu je stále otázkou budoucnosti. Odborníci se snažili o vytvoření alespoň poloautomatického rozpoznávače, který pracuje jen s určitým počtem předem zvolených proměnných. Jedním z takových systémů byl SAUSI (semi-automatic speaker identification system), který pracoval například s hodnotami f0 (Doherty & Hollien, 1978), dalším pak systém SASIS (Broderick, Paul & Rennick, 1975), jenž vyhodnocoval podobnost shodných fonetických událostí vybraných lidským operátorem. Ukázalo se, že automatické systémy bohužel nejsou dostatečně robustní na to, aby dokázaly odlišit inter-individuální variabilitu od variability způsobené např. různými nahrávacími podmínkami (Butcher, 2002, s. 7).

Výzkum možností využití suprasegmentálních jevů (např. základní frekvence, pauzy, trvání) ve forenzní fonetice se začíná objevovat později než výzkum segmentů, zhruba od 80. let 20. století. Důvodem může být nestálost těchto jevů, a tedy zhoršené možnosti porovnávání mluvčích zapříčiněné velkou intra-individuální variabilitou těchto faktorů. Příkladem je práce H. J. Künzela (1987), který na případu srovnávání f0 ukazuje, že všem zjištěným hodnotám je třeba přiřadit váhu s ohledem na obvyklost / jedinečnost jejich výskytu v populaci – čím periferněji (méně zastoupené) hodnoty, tím větší mají pro identifikaci mluvčího význam. Zároveň upozorňuje na to, že suprasegmentální prvky řeči mohou být zásadně ovlivněny např. stylem řeči mluvčího, jeho náladou či hlasitostí řeči.

V zájmu zachování co nejvyšší odborné úrovně, standardizace používaných postupů a sdílení poznatků byla v roce 1991 založena asociace forenzních fonetiků IAFPA (International Association for Forensic Phonetics and Acoustics) a posléze v roce 1995 společnost ENFSI (European Network of Forensic Science Institutes).

Dle programu konference IAFPA 2014 (Phonetisches Laboratorium (UZH), 2014) je v současné době pole výzkumu forenzní fonetiky široké, hodně se klade důraz na studium vlivu různých vnějších okolností na rysy hlasu, více se začínají zkoumat i suprasegmentální jevy jako rytmus či pauzy, zkoumá se maskování hlasu a jeho důsledky, spolehlivost laické identifikace, stálým tématem je využití formantů.

2 Úkoly forenzní fonetiky

Ačkoliv tvoří identifikace mluvčího největší část forenzní fonetiky, není to jediný úkol, který odborníci v této oblasti řeší. Přehled různých možných zadání uvádí Butcher (2002, s. 1–2):

1) identifikace mluvčího – nejčastější zadání
2) sporné vyjádření – identifikace obsahu promluvy v nahrávce, kdy jsou kvůli špatné kvalitě nahrávky či jiným okolnostem některé její části nesrozumitelné
3) autentifikace nahrávky – zjištění, zda bylo s nahrávkou nějakým způsobem manipulováno, např. vystřižením či záměnou pořadí některých částí a podobně
4) hlasové line-upy – konfrontace svědka (či oběti) trestného činu se sadou nahrávek, mezi nimiž je i hlas podezřelého

2.1 Předpoklady pro proveditelnost identifikace mluvčích

Zcela základním východiskem je fakt, že člověk je schopen rozpoznat jiného člověka mimo jiné i jen na základě poslechu jeho hlasu. Zatím nebylo dokázáno, že dva lidé nemohou mluvit naprosto stejným hlasem, ale možná variabilita je tak obrovská, že najít dva mluvčí, kteří se ve všech řečových aspektech shodují, by bylo velmi nepravděpodobné. Naproti tomu stojí fakt, že i jeden mluvčí může za různých okolností vykazovat různé hlasové charakteristiky. Úkolem forenzní fonetiky je tedy hledat takové rysy hlasu, jež jsou vůči různým podmínkám odolné, ale zároveň jsou i výrazně inter-individuální (to znamená, že se pro různé mluvčí liší), případně nějakým způsobem klasifikovat možné změny hlasu za různých podmínek, aby bylo možné porovnávat i nahrávky z různých prostředí (např. s různými ruchy v pozadí),

z různých kanálů (např. telefonní rozhovor × přímá nahrávka) nebo různý mluvní styl (např. křik × šepot). Výzkum zatím není v takovém stádiu, abychom byli schopni identifikovat mluvčího na 100 % – dosud nevíme, jakými všemi způsoby a do jaké míry je identita mluvčího v řeči zakódována.

Výsledná podoba řeči je ovlivněna souborem mnoha faktorů, z nichž některé pravděpodobně ještě ani neznáme (nebo nedokážeme určit celý jejich dosah). Pokud bychom ke srovnávání mluvčích použili jen jeden z těchto faktorů (např. základní frekvenci), dosáhneme tím pouze rozdělení populace do několika kategorií (např. nízká – průměrná – vysoká základní frekvence). Žádný z rysů není tak silný, abychom podle něj mohli bezpečně rozlišit všechny mluvčí – všichni mluvčí mají s někým dalším některé rysy společné, v jiných se ale naopak liší. Pokud ovšem použijeme kombinaci více rysů, zužuje se tím množina možných mluvčích, takže při dostatečném počtu zkoumaných faktorů je možné populaci mluvčích rozdělit na dostatečně malé kategorie, které jsou již ve forenzní praxi využitelné.

2.1.1 Inter-individuální variabilita řeči

Potencionální rozdíly mezi mluvčími můžeme rozdělit do dvou kategorií (např. Wolf, 1972, s. 2045):

- organické
- naučené

Organické rozdíly jsou zapříčiněny odlišnými rozměry a stavbou hlasivek a vokálního traktu (včetně různých vývojových vad či onemocnění) jednotlivých mluvčích. Patří mezi ně rychlost vibrace hlasivek (tedy základní frekvence hlasu, f0) a rezonantní frekvence hlasu (formanty). Za naučené rozdíly považujeme lingvistický systém, regionální a sociální vlivy na výslovnost. Obě kategorie spolu přitom souvisí – mohou ovlivňovat tytéž vlastnosti řeči (například výška formantů může být ovlivněna jak stavbou vokálního traktu, tak dialektem). Lze říci, že organické faktory určují krajní meze pro produkci řeči (např. člověk s dlouhými hlasivkami není schopen mluvit vysokým hlasem), naučené faktory pak určují, jakou část z fyzicky vymezeného prostoru budeme využívat a jakým způsobem.

2.1.2 Intra-individuální variabilita řeči

Jak je již zmíněno výše, hlas jedince není neměnný. Pro forenzní praxi je důležité, aby intra-individuální rozdíly řeči (tzn. u jednoho mluvčího) byly menší než rozdíly inter-individuální (tzn. mezi různými mluvčími). Po organických a naučených faktorech je zde tedy další, neméně důležitá skupina ovlivňující hlas, tentokrát bohužel v neprospěch identifikace mluvčích: aktuální psychické rozpoložení mluvčího, jeho fyzické zdraví, komunikační situace a záměr. Na případnou samotnou nahrávku pak dále mohou mít vliv kvalita záznamu, použité záznamové zařízení či komunikační kanál (typicky telefonní přenos) a okolní ruchy. Všechny tyto okolnosti způsobují, že hlas mluvčího může mít na různých nahrávkách různě pozměněné rysy. Nolan a Grigoras (2005, s. 168) si všímají, že supralaryngální parametry řeči mají menší intra-individuální variabilitu než parametry laryngální.

3 Proces identifikace mluvčího

Kromě samotného odborného srovnávání mluvčího, tedy porovnávání dvou nahrávek, existují v závislosti na výchozích podmínkách ještě další možné typy identifikace. Rozhodující je, zda existuje podezření na možného pachatele a zda disponujeme nahrávkou pachatelova hlasu, nebo se musíme spokojit se sluchovým svědectvím (viz tab. 3-1).

	K dispozici je nahrávka pachatele.	**Není k dispozici nahrávka, ale svědek (oběť) trestného činu**
Existuje podezření.	Srovnání mluvčích (Pokud podezřelý spolupracuje nebo máme k dispozici nějakou jeho starší nahrávku.)	Svědek zná podezřelého → svědek vydá prohlášení. Svědek nezná podezřelého → hlasová přehlídka (line-up).
Žádný podezřelý.	Profilace mluvčího a/nebo prezentace neznámého hlasu v médiích.	Odborníci se většinou nezapojují. Možná budoucnost: umělý zvukový obraz vzniklý hlasovou syntézou.

Tabulka 3-1: Přehled kombinací možných výchozích podmínek ve forenzní praxi (převzato z Jessen, 2010, s. 379).

Tabulka zároveň přehledně dělí identifikaci mluvčího na dva druhy dle odbornosti hodnotitele (Nolan, 1999):

- laická identifikace (vpravo) – svědek identifikuje mluvčího jen na základě poslechu

- technická identifikace (vlevo) – využití různých přístrojů a programů, statistické zpracování, ale i odborný poslech nahrávky (v rámci nějž lze zkoumat i obsah řeči, např. používání nezvyklých slov, nadužívání určitých parazitních výrazů, hezitace apod.)

3.1 Laická identifikace mluvčího

Byl-li svědkem (nebo dokonce obětí) trestného činu někdo, kdo pachatele neviděl (i třeba kvůli maskování), ale slyšel jeho hlas, přistupuje se v soudním řízení k tzv. laické identifikaci. Je otázkou, nakolik je takový postup obhajitelný, zvláště uběhl-li mezi spácháním trestného činu a soudním jednáním delší čas. V kapitole *Historický vývoj forenzní fonetiky* na str. 8 byly již zmíněny práce Francis McGehee (1937; 1944), v nichž prokázala snižující se schopnost správně si vybavit a rozpoznat slyšený hlas po delším čase.

Kromě časové prodlevy ovlivňuje rozpoznání mluvčího i fakt, zda svědek pachatele znal již z dřívější doby. Pokud ano, je třeba zjistit, do jaké míry je svědkovo tvrzení důvěryhodné – zda např. neslyšel hlas jen krátkou dobu, zda nebyl v pozadí slyšet i nějaký ruch apod. Z práce H. J. Künzela (1990, s. 35) totiž vyplývá, že lidé mají tendenci za zhoršených akustických podmínek považovat i odlišné hlasy za totožné. Obeznámenost s hlasem mluvčího dle Holliena et al. (1982) zvyšuje pravděpodobnost správného rozpoznání. Důležitá je ovšem délka promluvy – dle Rose a Duncan (1995) se úspěšnost rozpoznání mění v závislosti na trvání nahrávky od náhody až po téměř úplnou bezchybnost.

Jestliže svědek pachatele nezná, přistupuje se k tzv. hlasové přehlídce (lineupu). V podstatě jde o paralelu s vizuální identifikací pachatele v připravené skupině lidí (rekognicí). Pokud podezřelý spolupracuje, je vyhotovena nahrávka jeho hlasu, a to buď přečtením určitého textu – který může obsahovat stejná slova jako řeč pachatele, ale čtenost může mít na hlas podezřelého vliv (LTF hodnoty ve čtené a spontánní řeči porovnávala Moos, 2008), nebo nahrávkou spontánní řeči (která bude obsahově jiná než řeč pachatele, ale zase bude mít pravděpodobně blíže k původním okolním podmínkám řeči). Pakliže podezřelý nespolupracuje, může být podle právní úpravy daného státu možné použít alespoň nahrávku výslechu či starší nahrávky podezřelého, jsou-li k dispozici. Též musíme mít na zřeteli, že podezřelý

může na oko spolupracovat, ale při pořizování nahrávky se bude snažit měnit svůj hlas (více viz kapitola *Možné limitující faktory v identifikaci mluvčího*, s. 23).

Vytvořený hlasový line-up musí splňovat určitá pravidla, jinak může dojít ke snížení úspěšnosti identifikace (Eriksson, 2005, s. 12):

- Počet nahrávek: Zařazených mluvčích musí být přiměřený počet. Při malém počtu se může projevit efekt pořadí nahrávky. Pokud je hlasů naopak moc, úspěšnost rozpoznání klesá. Bull a Clifford (1984) udávají jako optimální asi 5–6 nahrávek.
- Podobnost nahrávek: Žádný hlas nesmí z ostatních nahrávek nějakým způsobem vyčnívat, aby jej nezaujatý posluchač kvůli tomu nevybral. Nežádoucí je ale i přílišná vzájemná podobnost nahrávek. Rothman (1977) dokládá pokles úspěšnosti rozpoznání z 94 % na 58 %, pokud srovnávané nahrávky pocházely od příbuzných (bratři, otec a syn).

Dále je třeba přihlédnout k tomu, jak „dobrý" je daný svědek v rozpoznávání mluvčích obecně. Pokud člověk dobře identifikuje známé hlasy, nemusí to ještě znamenat, že bude dobře identifikovat i hlasy cizí, neboť v obou případech je proces rozpoznávání odlišný (Eriksson, 2005, s. 11). Jako svědek či oběť trestného činu je člověk navíc vystaven stresu, což může mít na správné zapamatování hlasu vliv. Pokud rozpoznávají posluchači různé mluvčí v rámci experimentu, jsou na to předem připraveni. Míra úspěšnosti laického rozpoznání je tak reálně nižší, než uvádějí výsledky „laboratorních" výzkumů. V neposlední řadě je nutné počítat i s tím, že lidé mají tendenci přeceňovat své schopnosti rozeznat i cizí, jen jednou slyšený hlas (Nolan, 1999).

3.2 Technická identifikace mluvčího

3.2.1 Profilace mluvčího

Profilování mluvčího se používá v případech, kdy máme k dispozici nahrávku pachatele, ale zatím schází podezřelý. Právě pomocí profilace můžeme omezit skupinu možných pachatelů v ideálním případě až do té míry, že v kombinaci s dalšími výsledky vyšetřování zůstane jako možný pachatel jediný člověk.

Mediálně známým případem, v němž byla použita profilace mluvčího, byla kauza Yorkshirského Rozparovače, který v letech 1975–80 zavraždil 13 žen.

V průběhu vyšetřování byla pořízena nahrávka telefonátu, v němž se kdosi prohlašoval za Rozparovače. Tato nahrávka byla poskytnuta odborníku S. Ellisovi k vytvoření profilu mluvčího (Ellis, 1994). Ten především na základě analýzy dialektu velice přesně lokalizoval místo pobytu volajícího, zároveň však policii upozornil, že nahrávka může být podvržená. Vyšetřovatelé se této skutečně falešné stopy bohužel chytili a vyšetřování se nějaký čas ubíralo špatným směrem. Když byl pak následně volající identifikován, ukázalo se, že skutečně prožil celý život jen pár kilometrů od místa, které Ellis určil (French et al., 2006).

Profilací zařazujeme mluvčího do společenského rámce a zjišťujeme jeho individuální charakteristiky (což ale neznamená, že to jsou charakteristiky unikátní). Hodnotí se přitom vlastnosti hlasu z laického pohledu (výška a kvalita hlasu, mluvní tempo, případné výrazné prvky apod.). Na základě těchto charakteristik lze pak usuzovat jednak na fyzické parametry mluvčího, jednak na jeho regionální původ a sociální postavení (tab. 3-2).

výška těla	zdravotní stav	věk	pohlaví	sociolekt	regiolekt	cizí přízvuk (L2)	jazyk (L1)
←							→
organická / biologická (anatomie a fyziologie)				sociální			gramatická (lingvistický systém)

Tabulka 3-2: Domény ve forenzní klasifikaci mluvčích (převzato z Jessen, 2010, s. 381).

Jelikož základní frekvence hlasu i formantové frekvence souvisí s rozměry hlasivek, resp. vokálního traktu, lze předpokládat určitou míru korelace i s dalšími tělesnými rozměry, především s **výškou** (částečně i váhou). Tato korelace sice není nijak velká, obzvlášť ve středních hodnotách je velká variabilita, ale Greisbach (1999) alespoň došel k závěru, že mluvčí s velmi nízkými formanty (a tedy dlouhými hlasivkami a vokálním traktem) velmi pravděpodobně nebude malý a naopak mluvčí s vysokými formanty (a tedy krátkými hlasivkami a vokálním traktem) velmi pravděpodobně nebude vysoký. Tyto výsledky potvrdil i výzkum Jessena et al. (2005).

Z hlediska **zdravotního stavu** mluvčího jsou forenzně využitelné jen dlouhodobé jevy, například vývojové vady, různá trvalá onemocnění a poranění mluvního ústrojí, vývojové vady řeči či výslovnostní vady. Problematické bývá odhadnutí **věku**

mluvčího. Jak uvádí Jessen (2010, s. 383), podle statistiky BKA (Bundeskriminalamt) jsou nejčastějšími pachateli muži ve věku od 20 do 40 let. Nejvýraznější změny hlasu ale probíhají před tímto věkovým obdobím a po něm. Obecně lze říci, že u mužů asi do 40 let základní hlasová frekvence klesá a zhruba od 50 let začíná opět stoupat. U žen je patrný trvalý pokles f0. Ukazatelem věku může být i tempo řeči, které všeobecně s věkem klesá (Schötz, 2006). Věk mluvčího se většinou odhaduje na základě celkového dojmu z řeči. Rozdíl mezi odhadovaným a skutečným věkem bývá přibližně šest let, přičemž věk lépe odhadují forenzní odborníci než laici (Braun, 1996). Vodítkem může být i užívání neologismů (či naopak zastarávajících slov).

Určit **pohlaví mluvčího** bývá ve většině případů triviální záležitost. Zda je mluvčí muž, nebo žena, rozpoznává posluchač jednak podle základní frekvence, jednak podle vzájemného poměru vokálů. F0 se u mužů pohybuje přibližně v rozmezí 80–170 Hz (průměr 115 Hz), u žen je to asi 165–260 Hz (průměr 210 Hz) (Künzel, 1989). Je zde tedy určité pásmo, kde může hlas posluchači připadat příliš hluboký na ženu, ale i příliš vysoký na muže. V tom případě musí posoudit odborník, do jaké míry je hlas podobnější jedné z obou možností.

Na **regionální původ** a **společenský status** ukazují hlavně rysy jako volba slov, tedy obsah řeči (argot, žargon, typická nářeční slova nebo naopak užívání spisovných koncovek, složitější větné konstrukce), formantové frekvence vokálů (v moravské výslovnosti jsou vokály zavřenější) či jejich trvání (zkracování v ostravském nářečí). Výslovnost hlásek přitom může být ovlivněna nejen dialektem mluvčího (délka a zavřenost vokálů, asimilace znělosti), ale i jeho fyzickým stavem (závislost frekvence formantů na rozměrech vokálního traktu, různé druhy idiosynkratické výslovnosti jako třeba sigmatismus, rotacismus). Jak uvádí Campbell (1997), hodnocení segmentů je odolnější proti špatné kvalitě nahrávky.

3.2.2 Srovnávání mluvčích

Ke srovnávání nahrávek přistupujeme v momentě, kdy je dostupná nahrávka pachatele i podezřelého. Pokud to okolnosti umožňují, je vhodné, aby řeč podezřelého (srovnávací nahrávka) byla pořízena za stejných či alespoň podobných podmínek jako řeč pachatele (sporná nahrávka) – například šepot. Z nahrávek se poté případně vystříhají části, kde hovoří někdo jiný či které jsou pro srovnání nepoužitelné (např. velký hluk v pozadí). Je žádoucí, aby obě nahrávky byly co nejdelší (a obsahovaly tedy

co nejvíce porovnatelných rysů), v praxi se bohužel setkáváme i s nahrávkami trvajícími pouze několik sekund. Butcher (2002, s. 5) udává jako doporučené trvání nahrávky 15–120 s, v závislosti na zkoumaném jevu.

Nejčastějším zadáním bývá rozhodnout, nakolik je pravděpodobné, že obě nahrávky pocházejí od stejného mluvčího. Může se stát, že okruh podezřelých je již na základě předchozího vyšetřování definitivně stanoven (pokud jde například telefonát z určité kanceláře, jsou podezřelými pouze ti, kdo do ní mají přístup). V takovém případě je rozhodování snazší, je třeba pouze vybrat hlas, který se tomu na sporné nahrávce nejvíce podobá. Většinou ale není možné okruh podezřelých přesně definovat a pachatelem nemusí být nikdo z podezřelých.

3.2.2.1 Analýza dat

Samotné hodnocení a srovnávání nahrávek může probíhat dvěma způsoby (oba se uplatňují i v profilaci mluvčího):

- poslechová analýza
- akustická analýza

V obou případech se řeč redukuje na jednotlivé komponenty (např. f0, formanty, výslovnost hlásek apod.), jež se vyhodnocují samostatně. Nejlepšími rysy pro analýzu jsou ty, které vykazují co nejmenší intra-individuální variabilitu a naopak co největší variabilitu inter-individuální. V rámci **poslechové analýzy** hodnotíme na suprasegmentální úrovni základní frekvenci (intonaci), mluvní tempo, kvalitu hlasu, někdy i lexikální obsah, na segmentální úrovni pak výslovnost jednotlivých hlásek. Celkem má poslechová analýza čtyři hlavní části (Butcher, 2002, s. 4):

1) Hodnocení kvality hlasu (bez ohledu na další rezonance v nadhrtanových dutinách) – popis typu „napjatý hlas“, „dyšný hlas“, „třepená fonace“ apod.
2) Charakteristiky řeči, které vznikají ve vokálním traktu, artikulační nastavení. Patří sem dlouhodobé nastavení hrtanu, jazyka a rtů, rezonance měkkého patra. Můžeme zjistit např. přítomnou hypernazalitu či labializaci.
3) Aspekty artikulace, které vypovídají o regionálním a sociálním zázemí mluvčího. Především dialekty, ale i sociolekty, různé žargony či argot.

4) Identifikace idiosynkratické výslovnosti a anomálií. Zkoumají se především konsonanty a jejich výslovnost, např. lambdacismus, rotacismus, sigmatismus, ale i různé dysfluence. Často se využívá fonetická transkripce.

Akustická analýza umožňuje hodnotit a srovnávat i ty rysy řeči, které sluchem nevnímáme, většinou proto, že jako posluchači máme sklony ke generalizaci slyšeného, což umožňuje porozumění. (Pokud například dva různí mluvčí vysloví [a], vnímáme ho stále stejně, ačkoliv mluvčí mají rozdílné frekvence formantů.) Někteří zastánci poslechové analýzy akustickou analýzu odmítají s odůvodněním, že stále nemáme ucelené znalosti o tom, do jaké míry a jakým způsobem jsou získané hodnoty individuální, a že interpretace výsledků je možná jen v porovnání s korpusem hodnot obvyklých v populaci (zda jsou hodnoty mluvčího průměrné, nebo spíše vzácné) a takových korpusů existuje zatím pouze minimální množství. Výhodou akustické analýzy oproti poslechové je fakt, že výsledky lze zpracovat statisticky a kvantitativně. V současnosti je nejrozšířenější kombinace obou metod, tedy poslechově-akustická analýza, neboť každá metoda zkoumá nahrávku z jiného pohledu a obě se navzájem doplňují. Kombinovaný přístup používá např. německý BKA (např. Künzel, 1987).

Pomocí akustické analýzy hodnotíme obecně fyzikální vlastnosti řeči, tedy frekvenci, trvání a amplitudu. Jednotlivými sledovanými jevy pak může být základní frekvence hlasu, frekvence formantů, trvání jednotlivých hlásek či pauz, četnost pauz, modifikace hodnot za různých akustických podmínek apod.

Základní frekvence hlasu (f0, frekvence kmitání hlasivkových vazů) souvisí s percipovanou výškou hlasu. Průměrná f0 závisí jednak na pohlaví mluvčího a délce jeho hlasivek (inter-individuální variabilita), jednak může být ovlivněna i mluvním stylem (intra-individuální variabilita) – např. při křiku je f0 vyšší než při klidné řeči. Problémem je právě relativně velká intra-individuální variabilita (Braun, 1995), proto jsou potřeba výzkumy zjišťující, za jakých okolností a do jaké míry se průměrná f0 jednotlivých mluvčích mění (průměrnou f0 v několika mluvních stylech zkoumali např. Jessen et al., 2005). Hirson et al. (1995) zjistili, že lidé mají tendenci mluvit do telefonu víc nahlas, čímž se jejich průměrná základní frekvence zvyšuje. Průměrná f0 je jedním z mála rysů, pro nějž je k dispozici populační statistika obvyklosti hodnot (Hudson et al,. 2007). Kromě průměrné f0 lze zkoumat i změny základní frekvence v průběhu promluvy, např. intonační vzorce typické pro daného mluvčího.

Dalším dobrým individuálním znakem mluvčího jsou frekvence **formantů** (koncentrace akustické energie kolem určité frekvence u sonorních segmentů, způsobená rezonancemi vokálního traktu). Běžně se k analýze používají první tři formanty (F1–F3). Frekvence a vzájemná poloha F1 a F2 především rozlišuje jednotlivé druhy hlásek, F3 bývá označován za nejvíce individuální formant. Hodnoty je možné odečítat jednak ve středu hlásky, jednak v určitých intervalech po celou dobu jejího trvání (neboť frekvence formantů se může vlivem okolních hlásek v průběhu segmentu měnit). Tradičně se zkoumají hodnoty F1 a F2 pro určitou hlásku a jejich možná inter-individuální variabilita. Např. Nolan (1983) takto zkoumal střední hodnoty formantů u [r] a [l] v různých vokalických kontextech. Použil také tzv. F ratio – variabilita všech inter-individuálních středních formantových hodnot se vydělí variabilitou intra-individuální. Čím vyšší je F ratio, tím je hodnota vhodnější pro identifikaci mluvčích.

Potenciál tzv. dynamického způsobu analýzy formantů představila Goldstein (1976). Jedná se o výše zmíněný postup výpočtu formantových trajektorií měřením v celém trvání vokálu. McDougall (2006) zkoumá rysy diftongů a r-ových hlásek v různých hláskových kontextech a potvrzuje předpoklad, že dynamické změny formantů mezi akustickými cíli nesou větší inter-individuální variabilitu. Na tuto práci navázali např. Skarnitzl et al. (2012) výzkumem formantových trajektorií českých vokálů v různorodějších hláskových okolích – ta jsou pravděpodobně příčinou o něco menší zjištěné úspěšnosti identifikace. Vaňková (2013) pak porovnává výhodnost využití statických versus dynamických rysů formantů. Ačkoliv mírně lepší výsledky podle ní vykazují dynamické rysy, jako nejslibnější hodnotí statický formant F4.

Dlouhodobé spektrum (LTS, long-time spectrum) je poměrně spolehlivá metoda vyjádření průměrného rozložení energie ve frekvenční oblasti, využívaná např. v automatickém identifikačním systému SAUSI (Hollien, 1990). Pro rozpoznávání mluvčích není bohužel zcela vhodné, protože je ovlivněno i ruchy v pozadí nahrávky.

3.2.2.2 Možné limitující faktory v identifikaci mluvčího

Faktorů, které mohou ovlivnit úspěšnost identifikace mluvčího, je bohužel celá řada (Eriksson, 2005, s. 6–11). Zaprvé to jsou jevy přímo přítomné v hlasu mluvčího – např. záměrné maskování hlasu, cizí jazyk, obecná nestálost některých rysů (ať už způsobená fyzickým či psychickým stavem jedince, nebo jeho komunikačním

záměrem), dlouhý časový odstup mezi pořízením obou porovnávaných nahrávek. Zadruhé jsou to aspekty technického rázu, ovlivňující kvalitu nahrávky a možné vyhodnocení rysů řeči – krátké trvání nahrávky, šum nebo ruchy v pozadí, obecně špatná akustická kvalita nahrávky (způsobená nevhodným nahrávacím zařízením či přenosovými filtry).

Dlouhý časový odstup mezi nahrávkami nepatří mezi často zpracovávaná témata, poněvadž vyžaduje longitudinální studii. Hollien a Schwartz (2000; 2001) zkoumali nahrávky pořízené s odstupem 4 týdny až 20 let. Pozorovali přitom pokles úspěšnosti identifikace z 95 % při souběžných nahrávkách přes 70–85 % při odstupu nahrávek 4 týdny až 6 let až k úspěšnosti pouhých 35 % pro odstup 20 let (úspěšnost odborníků ale zůstala na 76 %). Zdá se tedy, že odstup mezi nahrávkami by neměl mít na úspěšnost identifikace příliš velký vliv, pokud nahrávky posuzují odborníci.

Pokusy o **maskování hlasu** kupodivu nejsou nijak výrazně rozšířené, Künzel (2000) uvádí, že se v nějaké formě vyskytují jen v asi 15–25 % případů, které BKA řešil za posledních 20 let. V posledních letech se ovšem začíná objevovat i elektronická manipulace hlasu nebo komunikace přes počítačovou hlasovou syntézu, což identifikaci mluvčího zcela vylučuje. Běžně používané metody jsou ale dle údajů BKA stále dost triviální – k maskování hlasu se nejčastěji používá falzet, třepená fonace, šepot, simulace cizího přízvuku nebo zacpávání nosu. Reich a Duke (1979) zkoumali úspěšnost identifikace mluvčího při použití různých druhů maskování hlasu. Největší vliv se ukázala mít hypernazalita, ostatní metody ovlivňovaly úspěšnost podobnou měrou. Orchard a Yarmey (1995) přišli se zjištěním, že srovnávání šepotu s fonovanými nahrávkami výrazně snižuje úspěšnost identifikace, lepších výsledků lze dosáhnout, pokud jsou šeptané obě srovnávané nahrávky.

Co se týče **cizího jazyka** nebo přízvuku v řeči mluvčího, výsledky studií nejsou jednoznačné. Podle již výše zmíněné práce McGehee (1937) není v úspěšnosti rozpoznávání mluvčích bez přízvuku nebo s cizím přízvukem rozdíl. Goldstein et al. (1981) se domnívají, že pokud má posluchač k dispozici dostatečně dlouhou nahrávku s cizím přízvukem, není rozdíl v úspěšnosti tak velký. Ovšem podle Dotyho (1998) poznávají posluchači mnohem lépe mluvčí bez cizího přízvuku. Podobné výsledky přináší i studie Thompsona (1987) či Schillera a Köstera (1996) – se silnějším cizím přízvukem (nebo dokonce cizí řečí) úspěšnost rozpoznání klesá. Stejně tak koreluje identifikace cizojazyčného mluvčího s mírou znalosti dotyčného jazyka u posluchače.

Častým technickým problémem je **krátké trvání** sporné nahrávky, běžně se setkáváme s promluvami dlouhými pouhých několik sekund. Pokud provádíme akustickou analýzu některého rysu řeči, je třeba zjistit, jaké minimální trvání musí v daném případě nahrávka mít (tzn. od jakého trvání jsou již hodnoty víceméně konstantní), a jestliže toto minimální trvání nahrávka nesplňuje, musíme výsledky brát s rezervou. Pojem „trvání nahrávky" zde přitom jde ruku v ruce s počtem měření či zvýšenou variabilitou jevu (Bricker & Pruzansky, 1966). Jak ovšem upozorňuje Yarmey (1991), delší nahrávky sice znamenají zvýšení úspěšnosti identifikace, nesou s sebou ale i zvýšený výskyt falešně pozitivních výsledků (kdy jsou za totožného označeni dva různí mluvčí). Je tedy třeba najít ideální trvání, kde bude počet falešně pozitivních výsledků redukován na minimum, ale úspěšnost identifikace již bude na použitelné úrovni.

Zhoršená akustická kvalita nahrávky komplikuje především akustickou analýzu. Temporální jevy (např. trvání segmentů) jsou přitom ovlivněny méně než jevy frekvenční (Foulkes & French, 2012). Nejběžnějšími komplikacemi je šum v pozadí nahrávky a vliv přenosových filtrů (telefonický přenos). S ústupem analogových technologií postupně mizí problémy způsobené převodem analogového signálu do digitální podoby, stále je třeba ovšem dbát na dostatečnou kvalitu nahrávky (vzorkování a kvantizace). Velká pozornost se v posledních letech věnovala vlivu telefonického filtrového pásma (300–3 400 Hz) na posun frekvence formantů. Vzhledem k vysoké spodní hranici filtru nelze z takovéto nahrávky extrahovat základní frekvence. Künzel (2001) zjistil posun F1 směrem k vyšším frekvencím (asi o 6 %) a na základě toho formantovou analýzu v telefonních nahrávkách nedoporučuje. Nolan (2002) oponuje, že lze porovnávat alespoň F2, jež by neměl být filtrem zasažen. Podle Künzela (2002) jsou ale problémem i další proměnné, např. vyšší hlasitost řeči při telefonování.

3.2.2.3 Vyvození závěrů

Jak již bylo zmíněno, data získaná pomocí akustické analýzy je třeba konfrontovat s korpusem hodnot pro celou populaci, aby bylo možné konstatovat, zda jsou naměřené hodnoty běžné, či spíše vzácné, a jaká je tak pravděpodobnost, že stejné hodnoty může vykazovat i jiný mluvčí. Podle Roberstona a Vignauxe (1995) by se výsledky měly vyjadřovat pomocí bayesovské statistiky. Obvyklým způsobem vyjádření

pravděpodobnosti je tzv. likelihood ratio (míra shody výsledků dělená obvyklostí výsledku v populaci), jehož správné použití ve své přednášce velmi názorně vysvětluje Morrison (2011). Pro většinu zkoumaných rysů bohužel dosud potřebné korpusy nebyly vytvořeny. Obecně lze říci, že negativní závěr (mluvčí jsou rozdílní) lze učinit s větší jistotou než závěr pozitivní (Nolan, 2001) – prakticky nikdy nemůžeme vyloučit, že dva mluvčí nemohou vykazovat stejné hodnoty. Panují-li mezi zjištěnými nahrávkami podstatné rozdíly, můžeme naopak totožného mluvčího prakticky bezpečně vyloučit (samozřejmě po zvážení možné intra-individuální variability a vnějších podmínek nahrávky). Dle IAFPA by závěr měl být vyjádřen na škále jistoty, do jaké míry se může jednat o shodného mluvčího.

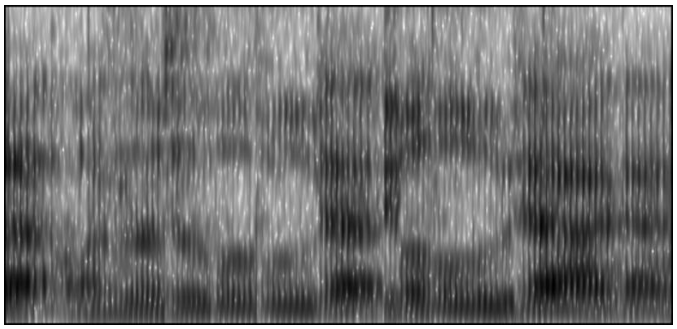
3.3 Metoda LTF

Porovnávání dlouhodobých formantových distribucí mluvčích (LTF, long-time formant distribution) je jedna z metod akustické analýzy řeči v rámci identifikace mluvčích. Jedná se o teprve krátce zkoumaný postup – navrhl jej Grigoras v roce 2005, v případu hledání autora obscénních telefonátů (Nolan & Grigoras, 2005). Zatímco autor původní poslechové analýzy se spíše klonil k názoru, že podezřelý a pachatel jsou jedna a tatáž osoba, akustická analýza provedená Nolanem toto vyloučila (srovnával formantové hodnoty diftongů v nahrávkách) a následná dodatečná Grigorasova analýza pomocí metody LTF dala Nolanovi za pravdu.

Metoda LTF vychází z techniky měření formantů jednotlivých hlásek (kdy např. zkoumáme, jak se liší hodnoty F1 a F2 dvou mluvčích při výslovnosti hlásky [a]). Odlišuje se ale v tom, že formantové hodnoty extrahujeme pro všechny sonorní segmenty v celém trvání nahrávky dohromady, bez toho, abychom rozlišovali jednotlivé druhy hlásek (viz např. Moos, 2008, s. 3). To je zároveň největší výhodou postupu, neboť není nutné nahrávky několikrát postupně poslouchat a kategorizovat jednotlivé vokály (někdy je obtížné určit, zda vyslovený vokál je např. spíše [e], nebo už [ɪ]), což bývá velmi časově náročné. Metoda LTF si naopak klade za cíl co nejrychlejší a nejsnazší přípravu dat. Vhodné segmenty pro analýzu proto vybíráme pouze vizuálně ze spektrogramu na základě přítomné viditelné formantové struktury. Vybrané hlásky se dále nekategorizují, což umožňuje i analýzu cizích jazyků (jejichž fonetický a fonologický systém se často odlišuje). Odpadá tím i faktor hodnotitele, který může některé hlásky špatně klasifikovat. Nevýhodou metody je, že nezohledňuje různé

formantové hodnoty u různých mluvčích pro jednotlivé hlásky, není tak třeba možné podchytit nářečně otevřené vokály.

Pro analýzu LTF se používají pouze sonorní hlásky (t.j. hlásky s patrnou formantovou strukturou) kromě nazál. Ty totiž ve zkoumaných frekvenčních hladinách obsahují místo formantů antiformanty. Jmenovitě tedy v českém prostředí používáme vokály /a/, /e/, /i/, /o/, /u/ (samozřejmě včetně dlouhých a diftongů), likvidy /r/, /l/, glidy /j/ a vokalické hezitace. Ostatní segmenty se z nahrávky odstraní, čímž vznikne tzv. vokalický proud (viz obr. 3-1). Z něj se následně pomocí formant trackeru, založeného na LPC predikci, každých 10 ms extrahují formantové hodnoty (které se pak hodnotí pro každý formant zvlášť). Před vlastní extrakcí je možné formantové trajektorie ručně upravit ve vhodném fonetickém programu. Automatická detekce formantů totiž stále ještě nepracuje bezchybně a občas se chytá špatných hodnot.



Obrázek 3-1: Ukázka spektrogramu vokalického proudu se zřetelnou formantovou strukturou jednotlivých segmentů (spektrum 0–5 000 Hz).

Extrahované hodnoty je možné dále zpracovávat dvěma způsoby (Jessen & Becker, 2010):

- průměrná LTF hodnota
- modelování LTF distribuce

Průměrnou LTF hodnotu získáme z aritmetického průměru všech formantových hodnot daného mluvčího, a sice pro každý formant zvlášť. Tento postup zvolila ve své práci např. Moos (2008). Výhodou je, že pracujeme s přehledným malým množstvím dat – pro každého mluvčího s pouze cca. třemi výchozími hodnotami (v závislosti na počtu zpracovávaných formantů). Jedna průměrná hodnota pro každý formant je také lépe statisticky uchopitelná, pokud porovnáváme větší množství

mluvčích či různé mluvní styly, přenosové kanály apod. – můžeme pracovat se směrodatnými odchylkami, jednoduše zkoumat korelace atd. Při **modelování LTF distribuce** naopak stále pracujeme se všemi extrahovanými hodnotami, kterých mohou být v závislosti na počtu měření i tisíce pro každý formant a mluvčího. Výhodou tohoto postupu je velká názornost a lepší rozlišovací schopnost. Dva mluvčí totiž mohou vykazovat shodné či velmi podobné průměrné LTF hodnoty, ale při pohledu na jejich LTF distribuce můžeme zjistit, že mají zcela jiný tvar (viz praktická část této práce). Nejjednodušším způsobem modelování LTF distribuce jsou histogramy rozdělení extrahovaných LTF hodnot (pro každý formant zvlášť). Tento postup použili např. Nolan a Grigoras (2005) a budeme jej používat i my v této práci. Becker et al. (2008) pro modelaci použili gaussovské mixture modely (GMM).

Metodu LTF distribuce nelze při forenzní identifikaci mluvčího používat samotnou. Její potenciál leží v rozdělení množiny podezřelých do kategorií s podobnými LTF distribucemi. To umožňuje provádět následné časově náročnější analýzy už jen na vybrané skupině podezřelých, kteří vykazují podobné LTF distribuce jako pachatel. Nelze vyloučit, ba je dokonce pravděpodobné, že různí mluvčí mohou mít velmi podobné LTF distribuce. O něco méně pravděpodobné je, že budou mít velmi podobné LTF distribuce všech (tří) měřených formantů, ale vyloučit to stále nelze. Proto pokud zjistíme ve dvou nahrávkách zcela odlišné LTF distribuce, můžeme relativně bezpečně vyloučit, že by šlo o téhož mluvčího (jsou-li nahrávky dostatečně dlouhé a pořízené za stejných podmínek). Pokud jsou si ale LTF distribuce v obou nahrávkách podobné, nelze o případné shodné totožnosti obou mluvčích rozhodnout bez další podrobné analýzy pomocí jiné techniky.

Jessen a Becker (2010) představili na 2. Panamerickém akustickém shromáždění (ASA 2nd Pan-American/Iberian Meeting on Acoustics) výsledky práce německého BKA. Zkoumali souvislost LTF hodnot a tělesné výšky. Výsledky jsou obdobné jako u Greisbacha (1999) – je vidět nepřímá úměra mezi LTF2, resp. LTF3 a tělesnou výškou. Korelace ovšem není nijak silná, velká variabilita panuje zvlášť ve středních hodnotách. Dále prokázali, že metoda LTF je dostatečně robustní vůči zpracování různými fonetiky a že různé jazyky obsazují zhruba stejný LTF prostor. Výsledky z jednoho jazyka je tedy s opatrností možné aplikovat i na jazyk jiný.

Anja Moos (2008; 2010) se zaměřila na porovnávání průměrných LTF hodnot v různých mluvních stylech (spontánní a čtená řeč) při telefonním přenosu. U většiny mluvčích zjistila vyšší hodnoty LTF2 a LTF3 při čtené řeči (LTF1 není možné při

telefonním přenosu zkoumat, neboť leží částečně mimo pásmo filtru). Průměrné LTF hodnoty ve spontánní a čtené řeči jednoho mluvčího spolu až na zmíněný posun korelují. Tvary LTF distribuce při spontánní a čtené řeči se mohou u mluvčího drobně lišit, ale v zásadě si bývají podobné. Dalším důležitým přínosem práce je zjištění minimálního trvání vokalického proudu, jež je nutné pro získání relevantních dat (to znamená, že v nahrávkách s delším trváním se již průměrná LTF hodnota nijak výrazně nemění). Všechny extrahované LTF hodnoty rozdělila do „paketů“ s různým počtem hodnot. Poté vždy vypočetla směrodatnou odchylku všech paketů se stejným počtem hodnot a směrodatné odchylky porovnala. Paket, počínaje jímž mají všechny větší pakety již téměř neměnnou směrodatnou odchylku, je možné považovat za potřebné minimální trvání vokalického proudu. Zjištěné hodnoty jsou průměrně cca. 7 s pro LTF1, 8 s pro LTF2 a 6 s pro LTF3 spontánní řeči, pro čtenou řeč jsou limity ještě o něco nižší. Je ovšem třeba podotknout, že řada mluvčích se od průměrného trendu liší a prahové hodnoty je pro ně nutné určit individuálně. Za výhodu metody LTF označuje Moos fakt, že je nezávislá na f0, dialektu a mluvním tempu.

3.3.1 Cíl práce

V praktické části této práce se budeme zabývat právě představenou metodou LTF a jejím uplatněním ve forenzní praxi. Všechny rysy využívané pro identifikaci mluvčího by měly být co nejvíce inter-individuálně variabilní a naopak co nejvíce intra-individuálně stabilní. Prvním úkolem tedy bude zjištění intra- a inter-individuální variability dlouhodobé formantové distribuce. Pokusíme se nastínit rozložení různých LTF hodnot v populaci, jež by mělo umožnit klasifikaci významnosti zjištěných hodnot např. pomocí likelihood ratia. Ve forenzní praxi jsme často nuceni pracovat s nahrávkami se špatnou akustickou kvalitou, proto je velká část studie věnována vlivu různé síly a druhu šumu v pozadí nahrávky na extrahované LTF hodnoty.

Snažíme se odpovědět na otázky, zda a případně do jaké míry dochází vlivem šumu ke změně tvaru rozložení LTF hodnot a zda změna probíhá u všech mluvčích stejně. Ovlivňují různé druhy či síly šumu nahrávku různými způsoby? Které formanty jsou nejvíce zasaženy? Je možné porovnávat nahrávky s různým druhem či silou šumu, především s čistou nahrávkou bez šumu? Pokud jsou změny zapříčiněné šumem takového rázu, že srovnání s čistou nahrávkou znemožňují, je možné šum nějakým způsobem redukovat či kompenzovat? Dalším častým problémem při identifikaci

mluvčích je příliš krátké trvání mnoha zkoumaných nahrávek. V závěrečné části se proto pokusíme zjistit, jaké minimální trvání musí mít nahrávka, aby extrahované hodnoty LTF distribuce měly dostatečně vypovídající hodnotu.

II Praktická část

4 Metoda

4.1 Příprava dat

Součástí zadání práce bylo nejprve pořízení a příprava nahrávek pro samotnou analýzu. Nahrávky byly vytvořeny podle pravidel korpusu VASST (Variabilita skupin a stylů), jehož se pak staly součástí. Tento korpus vzniká na Fonetickém ústavu FF UK a má za cíl mapovat oblastní, věkové a sociální řečové varianty. Aby byl korpus využitelný pro různé druhy experimentů, sestává každá nahrávka z pěti různých částí – řízeného dotazníku, spontánního rozhovoru, popisu obrázku, čtení vět a čtení souvislého textu. Nahrávky se pořizovaly u respondentů doma kvůli zajištění co nejpřirozenějších podmínek. Během nahrávání jsme se snažili eliminovat všechny rušivé elementy, jako hluk z ulice, zvuk spotřebičů apod.

Pro účely tohoto výzkumu byly zpracovány nahrávky deseti mužů ve věku 25 až 33 let z různých oblastí České republiky (konkrétně z jižních, západních a severních Čech), z různých sociálních vrstev a s různým nejvyšším dosaženým vzděláním. Tuto kategorii jsme zvolili s ohledem na to, že pachateli trestných činů bývají velice často muži ve věku 20 až 40 let (ČSÚ, 2013; Jessen, 2010, s. 383). Všechny subjekty jsou rodilými mluvčími českého jazyka a nehovoří žádným výrazným nářečím. Tazatelkami v nahrávkách jsou studentky Fonetického ústavu. Jako nahrávací zařízení byl použit přenosný rekordér se stereo mikrofonem Edirol R-09. Nahrávky byly vzorkovány na 48 kHz s hloubkou 16 bitů, stereo, a uloženy do formátu wav.

Ze dvou kanálů stereo nahrávky byl ponechán pouze kanál s hlasitějším zvukem a převzorkován na 32 kHz. Dále bylo třeba manuálně ztišit případné hlasité neřečové zvuky. U příliš tichých nahrávek byla hlasitost normalizována. Z nahrávky byla vybrána pouze část s neřízeným rozhovorem.

V programu Praat (Boersma & Weenink, 2013) byly části s rozhovorem kvůli lepší zpracovatelnosti rozděleny na zhruba minutové úseky (± 15 s). Do prvních dvou

vrstev textgridů byla přepsána řeč subjektu, do dalších dvou vrstev řeč tazatelky. S ohledem na předpokládané budoucí zpracování v rámci vytvářeného korpusu byla řeč rozdělena na jednotlivé úseky oddělené pauzami (pauzy > 120 ms). První ze dvou vrstev daného mluvčího obsahuje ortografický přepis řeči, který se kvůli jednotnosti zápisu řídí pravidly ÚČNK – nepoužívá se interpunkce a velká písmena (výjimkou jsou vlastní jména), nespisovné varianty jsou přepisovány nejbližšími spisovnými dle Jazykové příručky ÚJČ AV (2015). Přitakání a hezitace se přepisuje jako *hm* či *em*, případná nesrozumitelná slova jsou nahrazena značkou {unint#}, kde # značí počet slabik. Podobně jsou značkou {sensi#} označena místa s odstraněnými citlivými údaji. Druhá vrstva byla vytvořena přímo pro účely tohoto výzkumu a obsahuje přesný přepis řeči (např. používání nespisovných koncovek, vynechávání hlásek, komolení slov atp. – ne tedy fonetický přepis). Důvodem je použití rozpoznávacího programu, který na základě přepsaného textu lokalizuje hlásky v řečovém proudu, což by při spisovném (a tedy ne zcela souhlasícím) přepisu nebylo možné.

Minutové úseky byly dále rozsegmentovány na jednotlivé mezipauzové úseky subjektů. Na tyto úseky byl pak použit výše zmíněný rozpoznávací program, vyvinutý na Fonetickém ústavu FF UK ve spolupráci s FEL ČVUT (Pollák, Volín & Skarnitzl, 2007). Následně bylo nutné automaticky vytvořené hranice hlásek a slov zkontrolovat a případně ručně upravit (dle Machač & Skarnitzl, 2009), aby přesně odpovídaly skutečnosti, a vyřadit poškozené části promluv (např. nějaký externí zvuk nebo překrývající se řeč obou mluvčích). LTF metoda je sice navržena tak, aby podobné časově náročné a pracné úpravy nebyly zapotřebí, my jsme ale data takto pečlivě připravili, aby výsledný vokalický proud skutečně obsahoval pouze správné sonorní hlásky.

Z připravených nahrávek byly extrahovány řetězce hlásek využitelných pro metodu LTF (viz kapitola *Metoda LTF*, s. 26) s celkovým trváním přibližně 120 s pro každého mluvčího (kromě mluvčího PVC, kde bylo kvůli krátkému trvání rozhovoru možné pořídit řetězec pouze o délce asi 69 s). Kromě této základní čisté verze byly vytvořeny tři další varianty nahrávky se šumem – nahrávka s tzv. kavárenským šumem v pozadí, nahrávka s hnědým šumem a nahrávka s bílým šumem; vždy ve dvou verzích, kdy nahrávka měla od přidaného šumu odstup (SNR) 3 dB, resp. 10 dB. (Průměrnou intenzitu manipulované nahrávky jsme zjistili v Praatu a podle ní následně upravili amplitudu daného šumu.) Kavárenský šum (tzv. babble) byl nahráván v autentickém prostředí pomocí přenosného rekordéru se stereo mikrofonem Edirol R-09 (totožné nastavení jako u nahrávek řeči). Hnědý a bílý šum byly vygenerovány

v programu CoolEdit Pro (Syntrilium Software Corporation, 2002) pomocí výchozího nastavení. Nahrávka řeči byla pak vložena do jednoho a nahrávka šumu do druhého kanálu stereo záznamu. Opětovným sloučením do jednoho mono kanálu jsme obě nahrávky spojili. Následně byly ze smíchaných nahrávek automaticky extrahovány formantové hodnoty výše uvedeným způsobem.

Za účelem porovnávání nahrávek s různým druhem šumu byly obdobně vytvořeny další varianty hnědého a bílého šumu s jiným nastavením (změna nastavení intenzity při generování šumu). Další nahrávky kavárenských šumů byly staženy z internetu (Sounddogs.com, 2014). Všechny šumy jsme poté různě kombinovali s čistými nahrávkami (či jejich polovinami). Zdůvodnění postupů a detailnější informace viz kapitola *Simulace šumu, s. 55*.

V rámci zjišťování minimálního možného trvání vokalického proudu byly nahrávky rozděleny na různě dlouhé úseky (1/2, 1/4 a 1/8, tedy cca. 60, 30 a 15 s), jejichž histogramy rozdělení LTF hodnot jsme následně mezi sebou porovnávali. Tento experiment byl prováděn pouze na čistých nahrávkách bez šumu. V tabulce 4-1 je uveden seznam všech druhů vytvořených nahrávek včetně jejich akustické specifikace a přiřazené zkratky, používané dále v textu a na obrázcích.

Z výsledných vokalických proudů byly každých 10 ms extrahovány formanty F1–F3. Bylo použito výchozí nastavení Praatu, pouze počet formantů jsme zvýšili na 5,5 a maximální frekvenci na 5 500 Hz (dle Skarnitzl, Vaňková & Bořil, 2014) kvůli lepšímu rozpoznání vysokých hodnot. Pro každého mluvčího jsme tak získali zhruba 12 000 hodnot pro každý formant. (Výjimkou je opět mluvčí PVC, pro nějž bylo extrahováno jen asi 6 900 hodnot pro každý formant.) Zde je třeba upozornit, že tyto automaticky vyextrahované formantové hodnoty nebyly nijak kontrolovány a upravovány. Je tedy pravděpodobné, že určitá část hodnot neodpovídá skutečným pozicím formantů. Manuální kontrola nebyla provedena záměrně proto, že LTF metoda má být jednoduchá a rychlá, chtěli jsme tedy zjistit, zda je metoda spolehlivá i bez manuálních zásahů do formantových kontur či kontroly extrahovaných hodnot. Vzhledem k tomu, že hodnot bylo vyextrahováno velké množství, se dá předpokládat, že zcela chybné hodnoty tvoří v celkovém objemu relativně malou část, kterou dostatečně kompenzuje velký počet správných dat.

Druh nahrávky	Trvání	Označení
základní čistá nahrávka	120 s	[#]-c
nahrávka s kavárenským šumem č. 1 (-10 dB)	120 s	[#]-b10
nahrávka s kavárenským šumem č. 1 (-3 dB)	120 s	[#]-b3
nahrávka s posunutým kavár. šumem č. 1 (-10 dB)	120 s	[#]-2b10
nahrávka s posunutým kavár. šumem č. 1 (-3 dB)	120 s	[#]-2b3
nahrávka s kavárenským šumem č. 2 (-10 dB)	120 s	[#]-3b10
nahrávka s kavárenským šumem č. 2 (-3 dB)	120 s	[#]-3b3
nahrávka s kavárenským šumem č. 3 (-10 dB)	120 s	[#]-4b10
nahrávka s extrahovaným kavár. šumem č. 1 (-10 dB)	120 s	[#]-Xb10
nahrávka s hnědým šumem č. 1 (-10 dB)	120 s	[#]-bn10
nahrávka s hnědým šumem č. 1 (-3 dB)	120 s	[#]-bn3
nahrávka s hnědým šumem č. 2 (-10 dB)	120 s	[#]-2bn10
nahrávka s hnědým šumem č. 2 (-3 dB)	120 s	[#]-2bn3
nahrávka s bílým šumem č. 1 (-16 dB)	120 s	[#]-wn16
nahrávka s bílým šumem č. 1 (-10 dB)	120 s	[#]-wn10
nahrávka s bílým šumem č. 1 (-3 dB)	120 s	[#]-wn3
nahrávka s bílým šumem č. 2 (-10 dB)	120 s	[#]-2wn10
nahrávka s bílým šumem č. 2 (-3 dB)	120 s	[#]-2wn3
1. polovina čisté nahrávky	60 s	[#]-cA
1. pol. čisté nahrávky překrytá kavár. šumem č. 3 (-10 dB)	60 s	[#]-cA+3b10
2. pol. nahrávky s kavárenským šumem č. 1 (-10 dB)	60 s	[#]-b10B
1. pol. čisté nahrávky překrytá 2. polovinou nahrávky s kavárenským šumem č. 1 (-10 dB)	60 s	[#]-cA+b10B
1. pol. čisté nahrávky překrytá 2. polovinou nahrávky mluvčího LST s kavárenským šumem č. 1 (-10 dB)	60 s	[#]-cA+LST-b10B
1. pol. čisté nahrávky překrytá –„– mluvčího OBR –„–	60 s	[#]-cA+OBR-b10B
1. pol. čisté nahrávky překrytá –„– mluvčího ZRB –„–	60 s	[#]-cA+ZRB-b10B
1. pol. čisté nahrávky překrytá 2. polovinou čisté nahrávky mluvčího ZRB (-10 dB)	60 s	[#]-cA+ZRB-cB
čistá nahrávka rozdělená na čtvrtiny	30 s	[#] 1/4
čistá nahrávka rozdělená na osminy	15 s	[#] 1/8

Tabulka 4-1: Přehled nahrávek a jejich označení. Zástupný znak [#] značí zkratku mluvčího, popř. LTF1–3.

4.2 Analýza dat

Nejprve je třeba zjistit, zda je LTF metoda skutečně forenzně využitelná, to znamená, že hodnoty jednoho mluvčího jsou stabilně stejné (či velmi podobné), zatímco hodnoty různých mluvčích se co nejvíce odlišují. Pro zjištění intra-individuální variablity (v tomto případě spíše stability) bylo třeba rozdělit datový soubor (každý formant zvlášť) každého mluvčího na dvě poloviny (což by znamenalo délku nahrávky cca. 60 s, tzn. přibližně 6 000 vzorků), které jsme následně porovnali vizuálně na histogramech rozdělení LTF hodnot a také statisticky pomocí Kolmogorovova-Smirnovova testu (neparametrický test, jímž je možné otestovat, zda dva datové subsety mají shodné rozdělení hodnot, a zda tedy pocházejí z téhož datového souboru – zde od jednoho mluvčího). K rozdělení dat na poloviny byly použity tři různé metody:

1) Časové rozdělení – porovnání první poloviny nahrávky s druhou polovinou.
2) Systematické rozdělení – porovnání lichých vzorků nahrávky se sudými vzorky.
3) Náhodné rozdělení – porovnání náhodně vybrané poloviny vzorků s druhou náhodně vybranou komplementární polovinou.

Inter-individuální variabilitu formantů je vhodné znázornit graficky, aby bylo možné ji jednoduše vizuálně zhodnotit. Datový soubor (každý formant zvlášť) každého mluvčího jsme tedy převedli do histogramu. Všechny histogramy vytvořené v této práci mají stejné měřítko, aby bylo možné jednotlivé grafy snadno porovnávat:

- LTF1: rozpětí 200–1 000 Hz, krok 20 Hz
- LTF2: rozpětí 500–2 500 Hz, krok 50 Hz
- LTF3: rozpětí 1 700–3 700 Hz, krok 50 Hz

Další možností komparace mluvčích je porovnání jejich středních LTF hodnot. Všechny mluvčí jsme tedy porovnali po vzoru práce McDougall et al. (2012) pomocí krabicového grafu a mediány jejich LTF2 a LTF3 hodnot také v bodovém grafu.

Histogramy jednotlivých formantů jsme využili i při porovnávání nahrávek s různým druhem a intenzitou šumu a při hledání minimálního trvání nahrávky. Pro prezentaci výsledků není bohužel formát histogramů při větším počtu zkoumaných

mluvčích příliš praktický, protože jednotlivé grafy zabírají mnoho místa (což je umocněno ještě tím, že pro každého mluvčího zpracováváme tři formanty). Proto v této práci prezentujeme většinou jen příklady, kompletní seznam histogramů lze nalézt v příloze.

Pro kompaktnější zobrazení výsledků jsme proto na některých místech zvolili formát datové tabulky. Výsledné hodnoty jsou zde prezentovány pomocí percentilového rozpětí (rozmezí, v němž se nachází středních 90 % všech extrahovaných hodnot[1]) a mediánu (který se kvůli občasným extrémním hodnotám a výraznému sešikmení histogramu ukázal jako přesnější ukazatel střední hodnoty než aritmetický průměr). Touto metodou bohužel není možné postihnout vícečetné vrcholy (medián zpravidla vychází mezi ně) ani špičatost, ale vidíme alespoň rozptyl hodnot na frekvenční škále a sešikmení rozdělení.

Pro určení obvyklosti hodnot v populaci (likelihood ratia) jsme z výše zmíněných tabulkových percentilových hodnot (pouze z čistých nahrávek) sestavili grafy rozdělení, na kterých je vidět, ve kterých oblastech frekvenční škály jsou hodnoty mluvčích nejvíce nahuštěné.

[1] Hodnoty na 5. a 95. percentilu jsou dále v textu pro zjednodušení označovány jako frekvenční minima či maxima“.

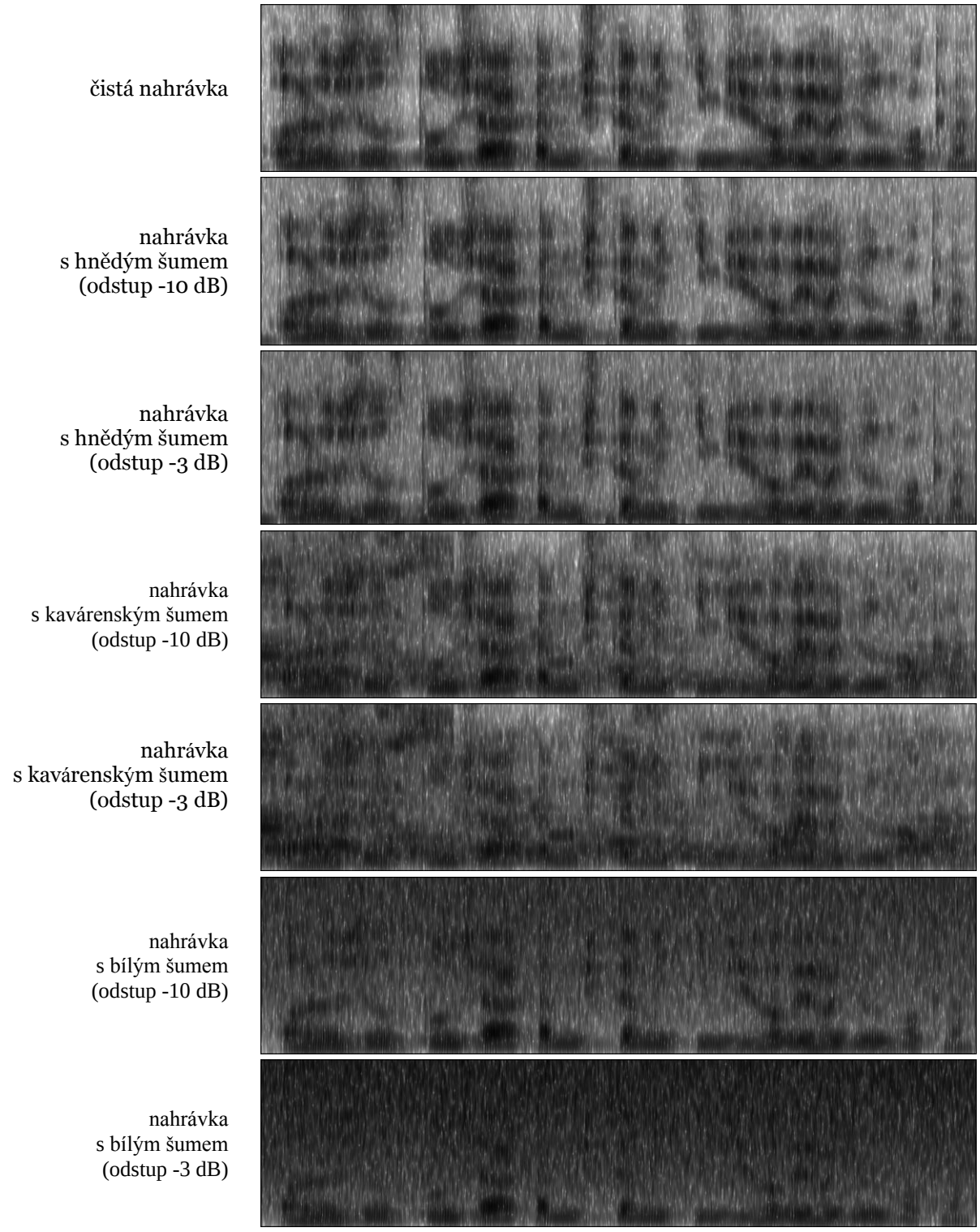
5 Výsledky a diskuze

5.1 Ovlivnění přípravy dat šumem v nahrávce

Šum v nahrávkách neovlivňuje jen automatickou extrakci formantů, ale má vliv už na manuální výběr hlásek vhodných pro LTF. Na následující straně jsou na obr. 5-1 zobrazeny výřezy ze spektrogramu téže části nahrávky mluvčího LST s různým typem a odstupem šumu. Je patrné, že hnědý šum ovlivňuje nahrávku téměř neznatelně, takže manuální výběr hlásek není nijak dotčen a omezen. Formantové struktury lze ještě relativně dobře rozeznat při zašumění kavárenským šumem s odstupem -10 dB. Při -3 dB se formantové struktury již poměrně dost ztrácejí, vybírat lze pouze ty nejvýraznější hlásky. To znamená, že abychom získali vokalický proud srovnatelného trvání jako u čisté nahrávky, musí být takto zašuměná vstupní nahrávka podstatně delší. Podobně je na tom i bílý šum s odstupem -10 dB. Nahrávka s bílým šumem o odstupu -3 dB je pak pro LTF zcela nepoužitelná. Vzhledem ke svému frekvenčnímu spektru a spektrálnímu sklonu, kdy ovlivňuje především vyšší frekvence, umožňuje bílý šum teoreticky alespoň analýzu f0, u velice výrazných vokálů i LTF1.

5.2 Předpoklady využití LTF hodnot ve forenzní praxi

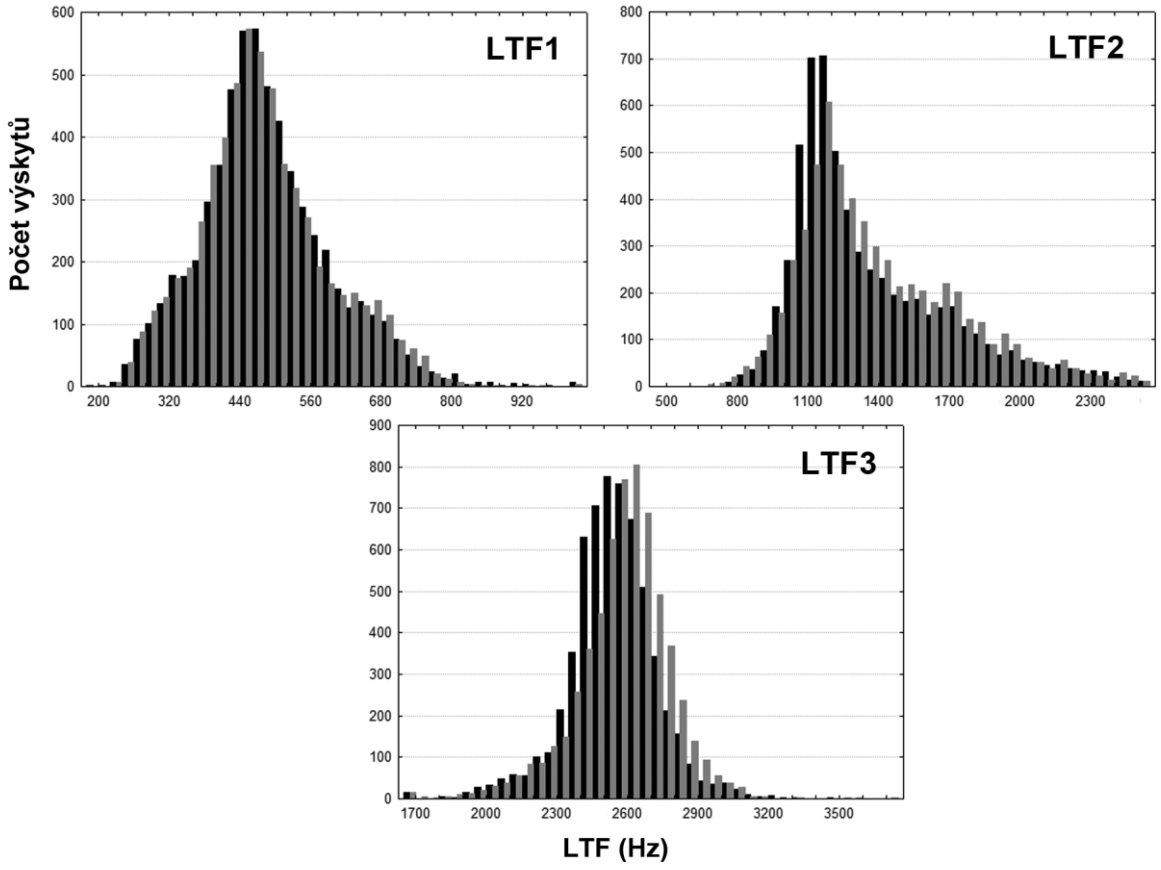
Stěžejním zjištěním, od něhož se odvíjí další postup výzkumu, je míra intra- a inter-individuální variability rozdělení LTF hodnot u jednotlivých mluvčích. Využití pro forenzní praxi vyžaduje, aby se hodnoty z různých nahrávek téhož mluvčího co nejvíce podobaly, a naopak nahrávky různých mluvčích aby se co nejvíce od sebe lišily.



Obrázek 5-1: Ukázky téže části spektrogramu pro čistou nahrávku a dále nahrávky s hnědým, kavárenským a bílým šumem, vždy v odstupu -10 a -3 dB (na frekvenční škále 0–5 000 Hz).

5.2.1 Intra-individuální variabilita

Rozdělením datových souborů jednotlivých mluvčích na poloviny a jejich následným porovnáním jsme zjišťovali, zda různé nahrávky od téhož mluvčího vykazují stejné rozdělení LTF hodnot (a tedy minimální intra-individuální variabilitu). Jak je vidět na obr. 5-2, vizuálně jsou rozdělení LTF hodnot v obou polovinách nahrávky téměř shodná už pro první typ dělení, tedy pro nahrávku rozdělenou časově. Bohužel statistické vyhodnocení pomocí Kolmogorovova-Smirnovova testu se ukázalo být při takovém objemu porovnávaných vzorků (každá polovina čítá cca. 6 tisíc měření) nepoužitelné – test je příliš citlivý a toleruje jen nepatrné odchylky. Pro všechny mluvčí vychází toto porovnání jako vysoce významné ($p < 0{,}001$).



Obrázek 5-2: Porovnání první (černá) a druhé (šedá) poloviny čisté nahrávky mluvčího MMS.

Proto jsme přistoupili k dalším způsobům dělení souboru, které by dle předpokladu měly zajistit lepší statistické výsledky – k dělení na náhodné poloviny a na sudé a liché vzorky. Pro náhodné poloviny už vysoce významný výsledek ($p < 0{,}001$) vychází pouze pro jednoho až dva mluvčí z deseti, u ostatních se pak aritmetický průměr LTF hodnot liší vždy jen o několik Hz. Při dělení na sudé a liché vzorky již Kolmogorovovův-Smirnovovův test vychází statisticky nevýznamně ($p > 0{,}1$) pro všechny mluvčí a aritmetické průměry LTF hodnot obou polovin se např. u F1 liší pouze o 0,3–1,4 Hz.

V dalším postupu práce jsme již tento test nevyužili, poněvadž bylo vždy již pohledem na porovnávané histogramy (a jejich odchylky) patrné, že by Kolmogorovovův-Smirnovovův test vyšel vysoce významně.

	KCR	LST	MMS	MSM	OBR	PNB	PTK	PVC	TZM	ZRB
F1 min.	309	281	314	233	309	260	298	310	266	295
	331	299	310	240	302	250	308	321	267	304
F1 max.	622	649	691	794	652	844	691	662	587	665
	629	650	689	740	658	799	717	669	581	677
F1 rozp.	313	368	377	561	343	584	393	352	321	370
	298	351	379	500	356	549	409	348	314	373
F1 med.	466	429	475	431	500	387	467	461	407	442
	474	441	468	429	496	386	476	454	399	438
F2 min.	859	965	997	920	868	877	1 001	802	847	919
	878	933	969	920	879	843	987	811	845	960
F2 max.	1 964	1 971	2 076	2 099	2 086	2 182	1 961	2 206	2 052	2 307
	1 997	1 872	2 059	2 122	2 165	2 222	2 028	2 274	2 081	2 293
F2 rozp.	1 105	1 006	1 079	1 179	1 218	1 305	960	1 404	1 205	1 388
	1 119	939	1 090	1 202	1 286	1 379	1 041	1 463	1 236	1 333
F2 med.	1 323	1 411	1 253	1 380	1 302	1 410	1 342	1 281	1 219	1 361
	1 330	1 332	1 309	1 344	1 352	1 391	1 337	1 384	1 270	1 393
F3 min.	1 953	2 272	2 222	2 194	2 083	2 239	2 245	2 175	2 108	2 129
	1 966	2 265	2 221	2 227	2 086	2 211	2 218	2 211	2 110	2 311
F3 max.	3 294	2 791	2 835	2 992	2 774	2 973	2 820	3 070	3 104	2 971
	3 348	2 796	2 876	2 936	2 842	2 939	2 869	3 052	3 023	2 929
F3 rozp.	1 341	519	613	798	691	734	575	895	996	842
	1 382	531	655	709	756	728	651	841	913	618
F3 med.	2 318	2 551	2 543	2 565	2 362	2 528	2 535	2 529	2 429	2 669
	2 309	2 569	2 597	2 548	2 397	2 518	2 544	2 556	2 434	2 670

Tabulka 5-1: Porovnání první a druhé poloviny nahrávky téhož mluvčího (uvedené vždy ve dvojicích pod sebou, v Hz).

V praxi musíme vycházet z časového dělení datových souborů, takže bohužel musíme konstatovat, že pro datové soubory čítající velký počet vzorků není Kolmogorovovův-Smirnovovův test prakticky použitelný, jelikož toleruje pouze odchylky menší, než jaké při porovnávání různých nahrávek (tím spíše u nahrávek se simulovaným

šumem, viz dále) běžně zaznamenáváme. Kladné výsledky testu pro rozdělení souborů na sudé a liché vzorky nicméně dokazují, že extrakční algoritmus formantů pracuje správně, neboť rozdíly mezi oběma polovinami nahrávek jsou zanedbatelné.

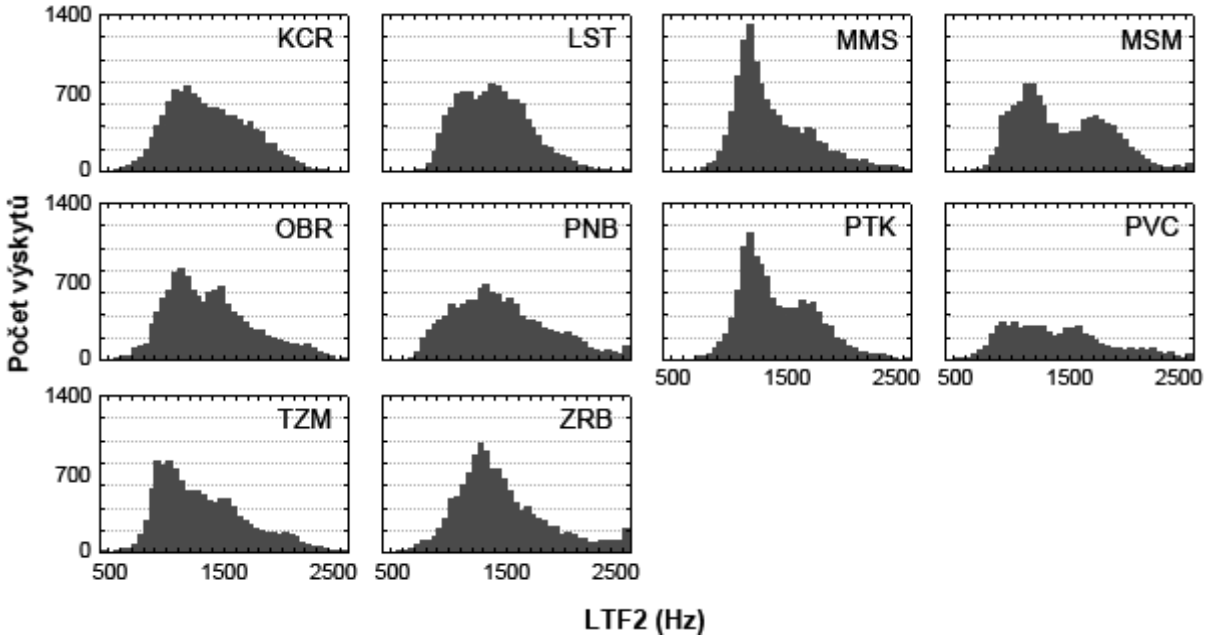
Podíváme-li se na číselné percentilové LTF hodnoty obou polovin nahrávek (tab. 5-1), můžeme konstatovat, že v případě F1 se hodnoty liší (až na dvě výjimky) o maximálně 20 Hz, pro F2 je rozdíl většinou nejvýše 50 Hz a u F3 taktéž až na několik výjimek do 50 Hz, což jsou (dle Kolmogorovova-Smirnovova testu) hodnoty možná statisticky významné, ovšem správnému vizuálnímu zhodnocení nijak výrazně nebrání.

5.2.2 Inter-individuální variabilita

Základem pro posuzování inter-individuální variability LTF hodnot byly histogramy jejich rozložení. Oproti jedinému číslu (průměru LTF hodnot), s nímž pracovala Moos (2008), mají velkou výhodu v tom, že na nich lze pozorovat a posuzovat jednak umístění vrcholu na frekvenční škále, jednak špičatost a sešikmení rozdělení, případně i výskyt vícečetných vrcholů, což vše může výrazně přispět k rozhodnutí o totožnosti mluvčího. Neméně důležitá je i jednoduchost hodnocení – odborník histogramy snadno vizuálně posoudí.

5.2.2.1 Čisté nahrávky

Na obrázku 5-3 je na příkladu porovnání LTF2 čistých nahrávek vidět výrazná inter-individuální variabilita v rozdělení LTF hodnot mluvčích. Totožnost mluvčího samozřejmě nelze určit na základě jediného takovéhoto porovnání – různí mluvčí mohou mít velice podobné histogramy. Můžeme se ale důvodně domnívat, že všechny mluvčí je možné rozdělit do určitých kategorií právě podle tvaru rozdělení LTF hodnot. To odpovídá předpokládanému využití analýzy LTF hodnot – rychlým a co nejméně pracným postupem nahrubo stanovit, zda porovnávané nahrávky vůbec mohou pocházet od jednoho mluvčího. Vždy je samozřejmě třeba brát v úvahu všechny zkoumané formanty dohromady, což výsledné kategorie zase dále diferencuje.



Obrázek 5-3: Porovnání histogramů hodnot LTF2 pro všechny mluvčí.

Pro LTF1 se frekvenční minima (hodnoty na 5. percentilu) jednotlivých mluvčích pohybují od 255 Hz do 319 Hz, tedy v poměrně úzké oblasti (viz tab. 5-2). Maxima (hodnoty na 95. percentilu) najdeme pak v rozmezí 585–767 Hz. (Nepočítaje v to mluvčího PNB, pro kterého automat chybně extrahoval větší množství vysokých hodnot, což bylo pravděpodobně způsobeno výskytem většího počtu nevýraznějších vokálů. Korigovaná horní hodnota by byla asi 650 Hz.) Minimální a maximální hodnoty spolu přitom nekorelují, což znamená, že každý mluvčí má individuální šířku frekvenčního spektra (percentilového rozpětí) – od 306 Hz po 530 Hz. Hodnoty mediánu vcelku přesně kopírují vrcholy histogramů (větší odchylka je jen u mluvčího MSM, jehož histogram má plochý vrchol a je silně pravostranně sešikmený). Vyskytují se v rozmezí 387 až 470 Hz.

Frekvenční minima LTF2 se nacházejí mezi 805–995 Hz, maxima v oblasti 1 930–2 303 Hz. Percentilové rozpětí je u druhého formantu 982–1 434 Hz a medián se pohybuje mezi 1 244 a 1 400 Hz. Medián zde vrcholu histogramu často neodpovídá, jelikož víceméně všichni mluvčí vykazují pravostranně sešikmené histogramy, asi polovina z nich má navíc histogramy bimodální (či alespoň s náznakem bimodality).

mluvčí	LTF1-c				LTF2-c				LTF3-c			
	min.	max.	rozp.	med.	min.	max.	rozp.	med.	min.	max.	rozp.	med.
KCR	319	625	306	470	867	1 977	1 110	1 327	1 961	3 321	1 360	2 313
LST	290	649	359	435	948	1 930	982	1 370	2 269	2 794	525	2 559
MMS	311	690	379	471	984	2 063	1 079	1 280	2 221	2 854	633	2 571
MSM	237	767	530	430	920	2 112	1 192	1 361	2 210	2 965	755	2 557
OBR	304	654	350	497	875	2 124	1 249	1 326	2 085	2 812	727	2 379
PNB	255	825 (650)	570 (395)	387	860	2 199	1 339	1 400	2 222	2 959	737	2 523
PTK	302	704	402	471	995	1 994	999	1 339	2 231	2 850	619	2 539
PVC	316	665	349	458	805	2 239	1 434	1 327	2 193	3 063	870	2 541
TZM	267	585	318	403	846	2 067	1 221	1 244	2 109	3 073	964	2 431
ZRB	300	669	369	439	930	2 303	1 373	1 376	2 213	2 954	741	2 669

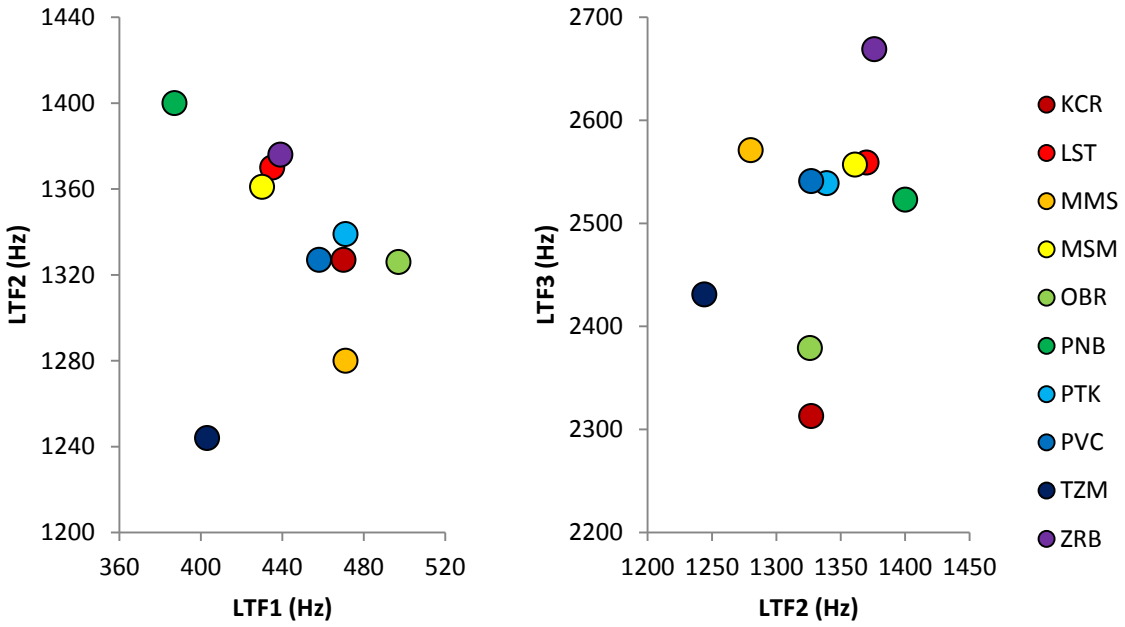
Tabulka 5-2: Hodnoty nacházející se na 5. a 95. percentilu v souboru extrahovaných hodnot, jejich percentilové rozpětí a mediány pro první, druhý a třetí formant čistých nahrávek (v Hz). Stínování naznačuje pozici na frekvenční škále vzhledem k ostatním mluvčím (čím tmavší barva, tím nižší hodnota).

Výše nastíněná inter-individuální variabilita hodnot LTF2 se u F3 opět částečně stírá. Již se zde nevyskytují vícečetné vrcholy a rozdělení hodnot různých mluvčích mají relativně podobný tvar. Výjimkou je zde mluvčí KCR, pro nějž automat vyextrahoval široké spektrum hodnot (1 961–3 321 Hz), takže jeho percentilové rozpětí je oproti jiným mluvčím až dvojnásobné. Důvodem je pravděpodobně dosti nedbalá výslovnost mluvčího KCR, jež se mimo jiné projevuje slabými vysokými formanty. Extrakční algoritmus se tedy velmi často chytá špatných hodnot. Hlavním srovnávacím parametrem pro LTF3 zůstává především pozice histogramu na frekvenční škále. Frekvenční minima a maxima ostatních mluvčích jsou zde 2 085–2 269 Hz, resp. 2 794–3 073 Hz. Percentilové rozpětí činí 525–964 Hz (což je o mnoho méně než u LTF2), mediány najdeme v hodnotách 2 379 až 2 669 Hz (šest mluvčích se přitom vejde do rozpětí pouhých 50 Hz).

Stínování v tabulce značí relativní pozici hodnoty na frekvenční stupnici vzhledem k hodnotám ostatních mluvčích. „Náhodné“ uspořádání různých odstínů zde naznačuje, že hodnoty frekvenčního maxima a minima (a tedy percentilový rozsah) na sobě globálně v populaci nezávisí, každý mluvčí je má nastaveny individuálně. Stejně tak spolu nesouvisí hodnoty formantů – např. mluvčí s nízkým LTF1 může mít nízké i další dva formanty (TZM), může je mít naopak vysoké (MSM) či třeba LTF2 nízký a LTF3 vysoký (PNB).

Jak vyplývá z předchozích odstavců, fakt, že jsme hlásky do vokalického proudu nevybírali vizuálně (podle výrazné formantové struktury), ale automaticky podle druhu hlásky, se podepsal na tom, že se do výběru dostaly i nedbale vyslovené hlásky s nevýraznými formanty. Citelně to u čistých nahrávek ovlivnilo pouze LTF1 u mluvčího PNB a LTF3 u mluvčího KCR. Naopak nám ale tento postup (automatický výběr hlásek do vokalického proudu a až následné přidání šumu) umožnil připravit pro všechny zašuměné nahrávky stejný výchozí soubor vokalického proudu obsahující stejné hlásky. Vzhledem k tomu, že intra-individuální variabilita hodnot těchto dvou mluvčí se nijak neliší od intra-individuální variability ostatních mluvčích, jsou chybně extrahované hodnoty rovnoměrně rozloženy po celém trvání nahrávek a mluvčí proto není nutné ze srovnávání vyřazovat.

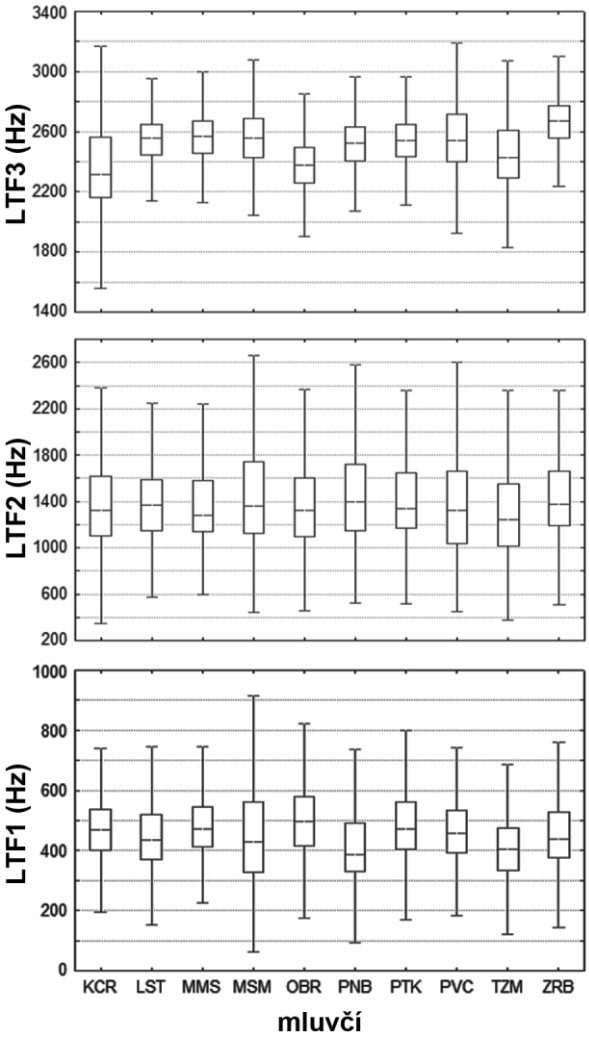
Na obr. 5-4 jsou v bodových grafech porovnány mediány LTF hodnot všech mluvčích pro závislosti LTF1×LTF2 a LTF2×LTF3. Můžeme si všimnout, že ačkoliv si jsou v tomto grafu mluvčí LST, MMS a ZRB podle mediánových hodnot LTF2 velmi podobní, histogramy jejich rozložení se viditelně liší (srov. obr. 5-3).



Obrázek 5-4: Srovnání mediánových LTF hodnot mluvčích.

Poslední zobrazovací / srovnávací metodou, kterou zde představíme, je krabicový graf (obr. 5-5). Velice dobře na něm vidíme, v jaké frekvenční hladině a rozpětí

se pohybují mediány LTF hodnot a jakou pozici a frekvenční šířku jednotlivá rozdělení mají. Z pozice mediánu na škále si můžeme udělat představu o sešikmení histogramu a z rozměrů středového boxu můžeme zhruba odečíst špičatost rozdělení. Tato metoda není ale schopna zachytit případné vícečetné vrcholy, které při identifikaci hrají velkou roli.

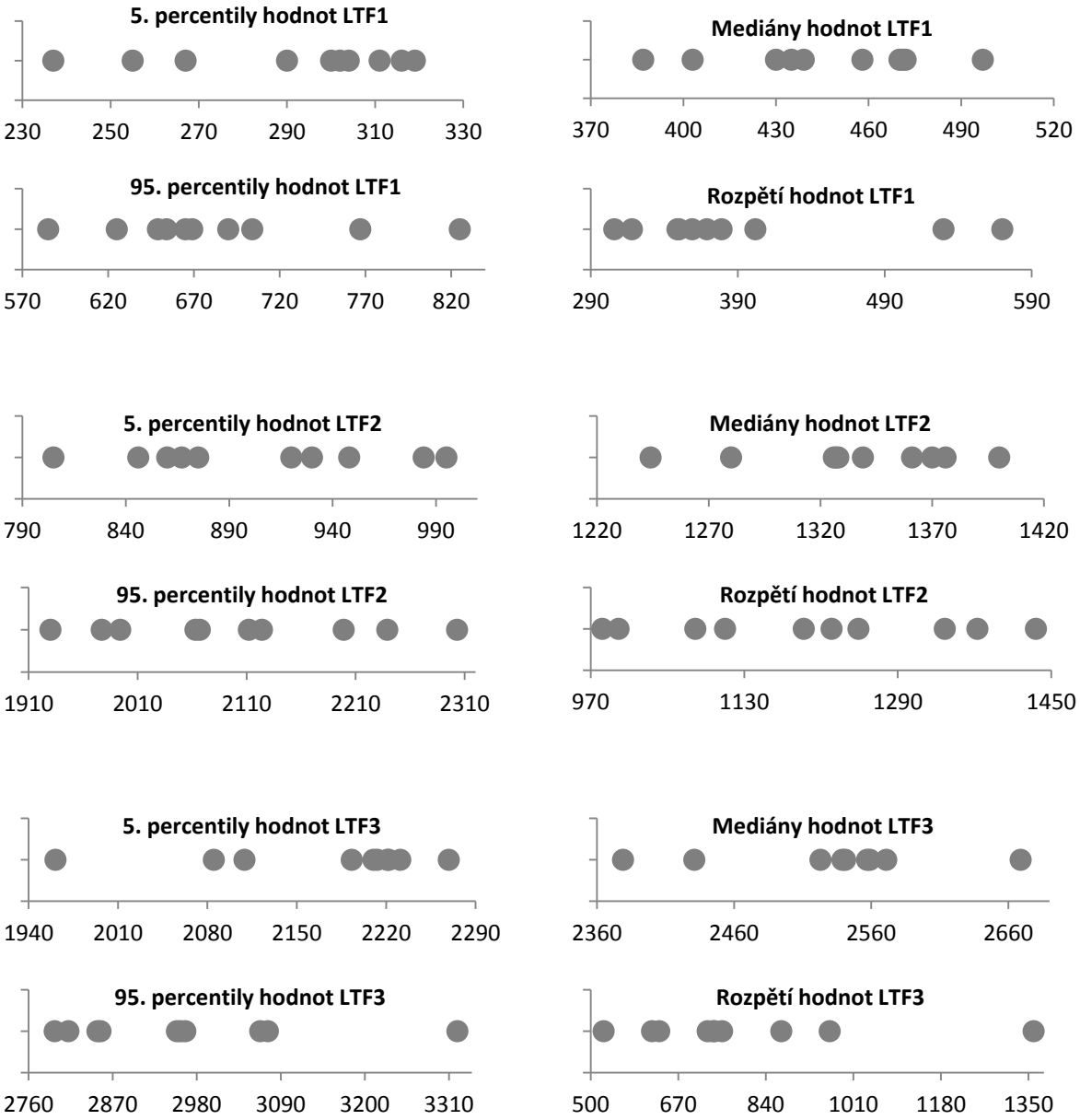


Obrázek 5-5: Srovnání rozdělení LTF hodnot všech mluvčích. Středová čára značí medián, krabice kvartilové rozpětí a fousky rozpětí bez odlehlých hodnot.

Obvyklost hodnot v populaci

Aby bylo možné určit, jak jsou naměřené hodnoty významné, je třeba zjistit, do jaké míry jsou dané hodnoty v populaci obvyklé (tzv. likelihood ratio). Mluvčí s unikátními hodnotami bude snadněji identifikovatelný než průměrný mluvčí. Pro vytvoření vypovídajícího korpusu by samozřejmě bylo zapotřebí získat data minimálně od desítek či stovek mluvčích, grafy v této práci mají tedy spíše jen ilustrační charakter.

Na níže uvedených grafech (obr. 5-6) vidíme porovnání tabulkových percentilových hodnot mluvčích. Lze předpokládat, že při dostatečně velké základně mluvčích budou mít LTF hodnoty tendenci tvořit normální rozdělení. Tomu relativně odpovídají všechny grafy až na minima (5. percentily) LTF1 a LTF3, která jsou poměrně výrazně levostranně sešikmená, a pravostranně sešikmené percentilové rozpětí (5–95. percentil) LTF1.



Obrázek 5-6: Zobrazení rozptylu percentilových LTF hodnot a mediánů ve skupině mluvčích (v Hz).

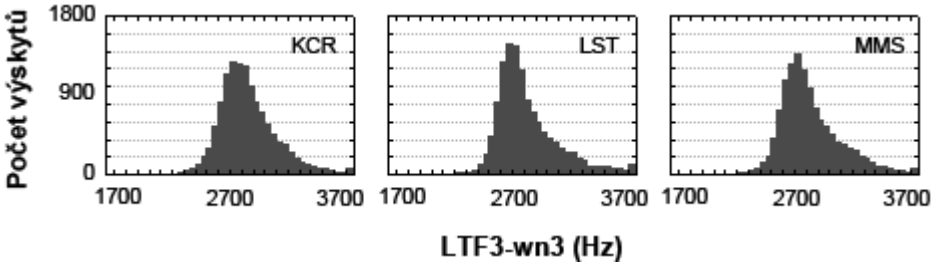
5.2.2.2 Nahrávky se šumem

Ukázalo se, že šum v nahrávce míru inter-individuální variability obecně snižuje, což je způsobeno tím, že extrakční algoritmus formantů reaguje kromě samotné řeči i na šum v pozadí. To vyextrahované hodnoty určitým způsobem zkresluje. Zjednodušeně můžeme říci, že pokud bychom nechali provést automatickou extrakci „formantů“ jen ze samotného šumu a z hodnot pak vytvořili histogram, budou se histogramy zašuměných nahrávek tomuto histogramu tvarově blížit, a rozdíly mezi jednotlivými mluvčími se tak budou stírat. K jak velkému zkreslení nahrávky dochází, je ovlivněno jednak odstupem šumu od nahrávky (čím menší odstup šumu, tím větší zkreslení a připodobnění „šumovému“ histogramu), jednak jeho druhem:

• hnědý šum – téměř žádné zkreslení → míra inter-individuální variability zůstává zachována

• kavárenský šum – u většiny mluvčích výrazné zkreslení, především v nižších formantech → míra inter-individuální variability klesá

• bílý šum – u všech mluvčích velice výrazné zkreslení → míra inter-individuální variability klesá, u F3 víceméně zaniká (viz obr. 5-7)

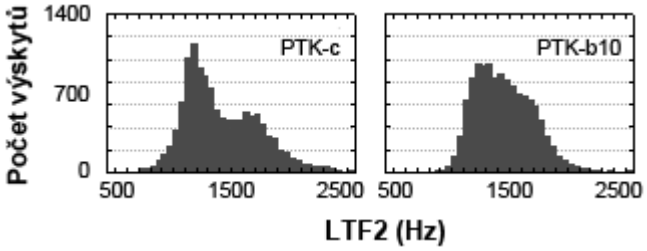


Obrázek 5-7: Ukázka zániku inter-individuální variability LTF3 hodnot při zašumění nahrávky bílý šumem s odstupem -3 dB.

5.3 Vliv různého odstupu a druhu šumu v nahrávce na extrahované LTF hodnoty

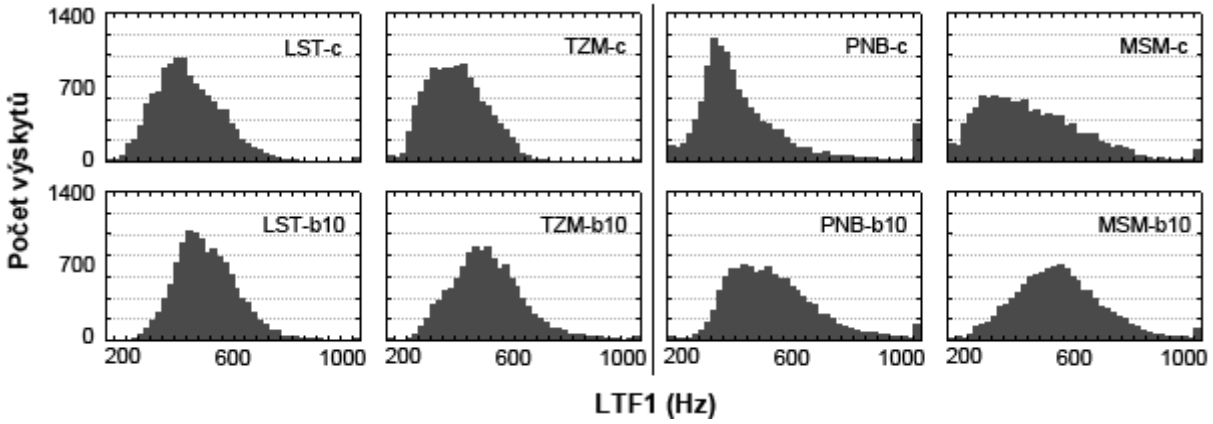
Jak jsme již naznačili v předchozí kapitole, extrakční algoritmus formantů produkuje z různě zašuměných nahrávek různé LTF hodnoty. Histogramy se pod vlivem kavárenského nebo bílého šumu připodobňují „šumovým“ histogramům (viz výše). To především znamená posun vrcholu grafu na frekvenční škále, ale také

změny v sešikmení a strmosti, a někdy dokonce i změnu počtu vrcholů grafu (typicky se dva vrcholy slučují do jednoho, viz obr. 5-8).



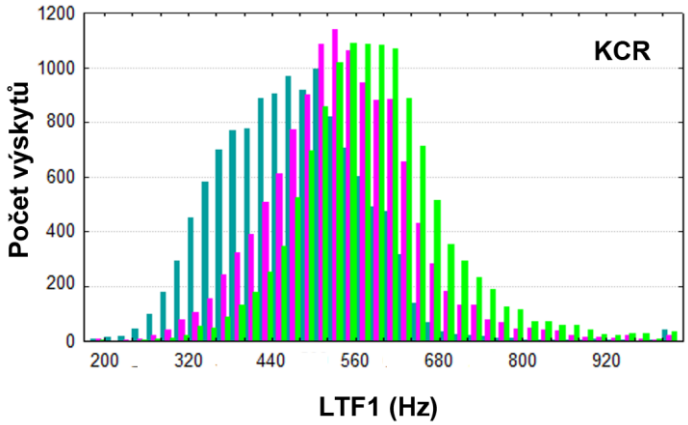
Obrázek 5-8: Ukázka změny počtu vrcholů vlivem šumu (čistá nahrávka versus kavárenský šum s odstupem -10 dB).

Zásadní překážkou pro forenzní praxi je fakt, že tyto změny nelze nijak zobecnit – u každého mluvčího a s každým druhem šumu probíhají v jiné míře a jiným způsobem (viz obr. 5-9). Jediným společným rysem je právě ono připodobňování k tvaru „šumového" histogramu, což je ve většině případů bez možnosti dalšího upřesnění bohužel irelevantní informace. Znamená to, že **nelze porovnávat nahrávky obsahující různý druh šumu.** (Za „druh šumu" je zde samozřejmě považováno i ticho, tedy čistá nahrávka.)



Obrázek 5-9: Porovnání chování nahrávek různých mluvčích po přidání šumu (kavárenský šum, -10 dB). Dvojice mluvčích vlevo a vpravo mají v čisté nahrávce podobné histogramy LTF1 hodnot. Zatímco levá dvojice zůstává relativně podobná i po přidání šumu (ačkoliv u mluvčího TZM je již náznak změny sešikmení), nahrávky mluvčích z pravé dvojice se chovají výrazně rozdílně.

Co se týče nahrávek s různým odstupem stejného druhu šumu, pak můžeme konstatovat, že tvar histogramu se u většiny mluvčích výrazně nemění, zpravidla dochází pouze k posunu histogramu na frekvenční škále (obr. 5-10).



Obrázek 5-10: Vliv kavárenského šumu s odstupem -10 dB (růžová) a -3 dB (zelená) na čistou nahrávku (modrá). Je patrné, že nejvíce se odlišuje čistá nahrávka, různé odstupy šumu už mají vliv víceméně jen na posun hodnot na frekvenční škále.

Teoreticky by mělo být možné využít alespoň posunu vrcholu na frekvenční stupnici: Dejme tomu, že máme porovnat dvě nahrávky, z nichž jedna je určitým způsobem zašuměná. Přitom čistá nahrávka vykazuje vrchol hodnot LTF1 např. kolem 350 Hz, zatímco zašuměná nahrávka má vrchol kolem 500 Hz. Ze zašuměné nahrávky se nám povedlo extrahovat část šumu (postup viz kapitola *Extrakce šumu z původní nahrávky*, s. 57), jehož pseudo-LTF1 hodnoty[2] se pohybují kolem 440 Hz. Vezmeme-li v úvahu, že nárůst LTF hodnot by měl být v pořadí „čistá nahrávka – zašuměná nahrávka – samotný šum“, můžeme vyloučit, že by obě nahrávky pocházely od jednoho mluvčího. V tom případě by se totiž vrchol hodnot LTF1 v zašuměné nahrávce musel nacházet v rozmezí 350–440 Hz.

Následují přehledové tabulky krajních a mediánových hodnot a hodnoty percentilového rozpětí. V každé jsou porovnávány nahrávky s oběma odstupy určitého šumu s čistou nahrávkou. Jak již bylo řečeno, nejméně mění rozložení LTF hodnot hnědý šum (tab. 5-3). Minimální hodnota LTF1 zůstává u obou odstupů stejná jako u čisté nahrávky, ale u některých mluvčích se zvyšuje (především u šumu -3 dB)

[2] Tzn. hodnoty, které by vyextrahoval formantový algoritmus. „Pseudo-“ proto, že v tomto případě se nejedná o vokalický proud.

hodnota maximální (a tedy i percentilové rozpětí). I mediány zůstávají téměř shodné, pouze u mluvčího MSM se medián posunuje asi o 20 Hz nahoru. Je třeba zmínit, že v nahrávkách s hnědým šumem s odstupem -3 dB výrazně vzrostl počet extrahovaných hodnot s frekvencí vyšší než 1 000 Hz (pravděpodobně tedy špatně rozpoznaných).

	druh nahrávky	KCR	LST	MMS	MSM	OBR	PNB	PTK	PVC	TZM	ZRB
F1 min.	c	319	290	311	237	304	255	302	316	267	300
	bn10	319	292	308	238	306	256	302	315	266	297
	bn3	311	291	301	239	307	254	301	313	266	293
F1 max.	c	625	649	690	767	654	825	704	665	585	669
	bn10	631	655	682	790	661	880	706	677	594	667
	bn3	665	678	681	964	685	1 021	726	714	594	685
F1 rozp.	c	306	359	379	530	350	570	402	349	318	369
	bn10	312	363	374	552	355	624	404	362	328	370
	bn3	354	387	380	725	378	767	425	401	328	392
F1 med.	c	470	435	471	430	497	387	471	458	403	439
	bn10	473	437	470	433	501	390	470	459	407	437
	bn3	480	440	469	451	509	399	472	464	407	439
F2 min.	c	867	948	984	920	875	860	995	805	846	930
	bn10	892	963	993	927	886	885	1 007	817	867	972
	bn3	935	994	1 008	945	911	925	1 038	843	867	1 033
F2 max.	c	1 977	1 930	2 063	2 112	2 124	2 199	1 994	2 239	2 067	2 303
	bn10	1 977	1 928	2 026	2 082	2 124	2 209	1 975	2 203	2 025	2 156
	bn3	2 012	1 936	1 987	2 159	2 136	2 245	1 972	2 142	2 025	2 076
F2 rozp.	c	1 110	982	1 079	1 192	1 249	1 339	999	1 434	1 221	1 373
	bn10	1 085	965	1 033	1 155	1 238	1 324	968	1 386	1 158	1 184
	bn3	1 077	942	979	1 214	1 225	1 320	934	1 299	1 158	1 043
F2 med.	c	1 327	1 370	1 280	1 361	1 326	1 400	1 339	1 327	1 244	1 376
	bn10	1 353	1 380	1 284	1 395	1 339	1 423	1 346	1 354	1 279	1 388
	bn3	1 419	1 403	1 298	1 477	1 375	1 476	1 377	1 421	1 279	1 425
F3 min.	c	1 961	2 269	2 221	2 210	2 085	2 222	2 231	2 193	2 109	2 213
	bn10	1 997	2 284	2 250	2 228	2 096	2 237	2 247	2 199	2 131	2 291
	bn3	2 053	2 298	2 281	2 248	2 118	2 254	2 268	2 216	2 131	2 338
F3 max.	c	3 321	2 794	2 854	2 965	2 812	2 959	2 850	3 063	3 073	2 954
	bn10	3 300	2 801	2 839	2 978	2 844	3 030	2 848	3 057	3 024	2 905
	bn3	3 272	2 828	2 834	3 156	2 945	3 174	2 892	3 078	3 024	2 907
F3 rozp.	c	1 360	525	633	755	727	737	619	870	964	741
	bn10	1 303	517	589	750	748	793	601	858	893	614
	bn3	1 219	530	553	908	827	920	624	862	893	569
F3 med.	c	2 313	2 559	2 571	2 557	2 379	2 523	2 539	2 541	2 431	2 669
	bn10	2 342	2 562	2 570	2 561	2 385	2 527	2 541	2 538	2 434	2 657
	bn3	2 397	2 568	2 565	2 583	2 397	2 537	2 549	2 539	2 434	2 646

Tabulka 5-3: Porovnání hodnot v čisté nahrávce a v nahrávce s hnědým šumem (-10 a -3 dB) pro první, druhý a třetí formant (v Hz).

Minimální hodnoty LTF2 a LTF3 v zašuměných nahrávkách rostou (zhruba o několik desítek Hz), maximální hodnoty naopak většinou klesají. Zajímavý je mluvčí MSM, u kterého maximum LTF2 u šumu -10 dB klesá, ale při -3 dB je dokonce výše než v čisté nahrávce. Medián si u všech mluvčích zachovává rostoucí tendenci (opět v desítkách Hz) či u některých mluvčích v LTF3 stagnuje.

	druh nahrávky	KCR	LST	MMS	MSM	OBR	PNB	PTK	PVC	TZM	ZRB
F1 min.	c	319	290	311	237	304	255	302	316	267	300
	b10	377	358	385	321	391	336	386	385	332	364
	b3	427	416	435	397	439	388	430	435	391	415
F1 max.	c	625	649	690	767	654	825	704	665	585	669
	b10	709	699	731	817	721	825	734	723	737	694
	b3	768	753	788	826	776	828	780	767	792	758
F1 rozp.	c	306	359	379	530	350	570	402	349	318	369
	b10	332	341	346	496	330	489	348	338	405	330
	b3	341	337	353	429	337	440	350	332	401	343
F1 med.	c	470	435	471	430	497	387	471	458	403	439
	b10	537	499	525	539	559	509	530	526	503	503
	b3	577	557	576	586	594	574	577	575	566	558
F2 min.	c	867	948	984	920	875	860	995	805	846	930
	b10	1 082	1 073	1 077	1 040	1 010	1 075	1 104	1 015	1 028	1 117
	b3	1 165	1 148	1 130	1 131	1 094	1 169	1 145	1 144	1 122	1 171
F2 max.	c	1 977	1 930	2 063	2 112	2 124	2 199	1 994	2 239	2 067	2 303
	b10	1 938	1 886	1 899	1 980	1 979	2 047	1 883	1 948	1 954	1 918
	b3	1 955	1 909	1 900	1 976	1 955	2 009	1 895	1 943	1 973	1 916
F2 rozp.	c	1 110	982	1 079	1 192	1 249	1 339	999	1 434	1 221	1 373
	b10	856	813	822	940	969	972	779	933	926	801
	b3	790	761	770	845	861	840	750	799	851	745
F2 med.	c	1 327	1 370	1 280	1 361	1 326	1 400	1 339	1 327	1 244	1 376
	b10	1 473	1 431	1 382	1 485	1 413	1 492	1 425	1 441	1 451	1 444
	b3	1 510	1 477	1 451	1 520	1 470	1 523	1 476	1 489	1 509	1 483
F3 min.	c	1 961	2 269	2 221	2 210	2 085	2 222	2 231	2 193	2 109	2 213
	b10	2 128	2 280	2 288	2 257	2 126	2 282	2 261	2 225	2 243	2 269
	b3	2 200	2 304	2 304	2 278	2 172	2 284	2 274	2 260	2 289	2 279
F3 max.	c	3 321	2 794	2 854	2 965	2 812	2 959	2 850	3 063	3 073	2 954
	b10	3 081	2 885	2 931	3 025	2 900	2 997	2 961	3 005	3 023	2 952
	b3	3 040	2 943	2 973	3 035	2 947	3 012	2 992	3 013	3 023	2 979
F3 rozp.	c	1 360	525	633	755	727	737	619	870	964	741
	b10	953	605	643	768	774	715	700	780	780	683
	b3	840	639	669	757	775	728	718	753	734	700
F3 med.	c	2 313	2 559	2 571	2 557	2 379	2 523	2 539	2 541	2 431	2 669
	b10	2 589	2 582	2 625	2 669	2 437	2 609	2 613	2 609	2 614	2 657
	b3	2 651	2 615	2 659	2 693	2 512	2 651	2 650	2 656	2 671	2 664

Tabulka 5-4: Porovnání hodnot v čisté nahrávce a v nahrávce s kavárenským šumem (-10 dB a -3 dB) pro první, druhý a třetí formant (v Hz).

Kavárenský šum již nahrávku mění výrazněji a extrahované hodnoty se často liší od původních (viz tab. 5-4). Např. minimum a maximum LTF1 se v šumu s odstupem -10 dB posunují zhruba o 40–100 Hz nahoru, u šumu s odstupem -3 dB pak o dalších cca. 40–90 Hz. Zvýšení frekvence pro minimální hodnoty u LTF2 se pohybuje už kolem 100–200 Hz (pro každý odstup zvlášť). Maximální hodnoty ovšem klesají, takže frekvenční percentilový rozsah se zužuje až o stovky Hz. U LTF3 už nejsou změny minima a maxima tak markantní (asi do 150 Hz), zajímavé ovšem je, že hodnota maxima u většiny mluvčích opět roste. Hodnoty mediánu se zvyšují u všech tří formantů (vždy pro šum s odstupem -10 dB + pro šum s odstupem -3 dB):

- F1: asi 50–100 Hz + dalších asi 50 Hz
- F2: asi 70–200 Hz + dalších asi 40–70 Hz
- F3 asi 30–100 Hz + dalších asi 10–70 Hz

Nejvíce destruktivní vliv má na nahrávku bílý šum (viz tab. 5-5). Minimální hodnoty LTF1 sice v zašuměných nahrávkách rostou jen v desítkách Hz, ale už medián LTF1 stoupá až o 300 Hz a maximální hodnota LTF1 roste při šumu -10 dB o 300–700 dB a při šumu -3 dB o dalších 100–300 Hz. Přitom zde platí nepřímá úměra, tedy čím větší je rozdíl mezi čistou nahrávkou a nahrávkou s bílým šumem -10 dB, tím menší rozdíl je mezi nahrávkami se šumem -10 dB a -3 dB. Minimální a maximální hodnoty LTF2 se zvyšují v řádu stovek Hz. U LTF3 je u minimálních a maximálních hodnot největší skok mezi čistou nahrávkou a přidáním šumu. Zesilování šumu zvyšuje extrahované hodnoty už podstatně méně. Mediány všech tří formantů opět pod vlivem šumu rostou (o stovky Hz).

Všechny šumy, jež jsme v rámci této práce testovali, posouvaly vrcholy rozdělení LTF hodnot směrem nahoru, do vyšších frekvencí. Nelze ale vyloučit, že s jiným druhem šumu (např. hluboký zvuk motoru) či u jiných mluvčích (např. ženský hlas) nemůže docházet i k posunu směrem k nižším frekvencím.

	druh nahrávky	KCR	LST	MMS	MSM	OBR	PNB	PTK	PVC	TZM	ZRB
F1 min.	c	319	290	311	237	304	255	302	316	267	300
	wn10	341	313	345	268	352	290	353	341	279	318
	wn3	399	365	426	355	422	334	412	379	339	372
F1 max.	c	625	649	690	767	654	825	704	665	585	669
	wn10	960	1 016	1 028	1 460	1 052	1 413	1 075	1 006	1 298	1 083
	wn3	1 283	1 395	1 300	1 518	1 315	1 508	1 312	1 293	1 484	1 364
F1 rozp.	c	306	359	379	530	350	570	402	349	318	369
	wn10	619	703	683	1 192	700	1 123	722	665	1 019	765
	wn3	884	1 030	874	1 163	893	1 174	900	914	1 145	992
F1 med.	c	470	435	471	430	497	387	471	458	403	439
	wn10	544	508	536	565	588	518	530	517	489	486
	wn3	640	609	642	722	664	713	629	590	611	586
F2 min.	c	867	948	984	920	875	860	995	805	846	930
	wn10	1 305	1 238	1 132	1 134	1 099	1 273	1 154	1 310	1 141	1 224
	wn3	1 555	1 483	1 300	1 436	1 321	1 517	1 343	1 600	1 459	1 380
F2 max.	c	1 977	1 930	2 063	2 112	2 124	2 199	1 994	2 239	2 067	2 303
	wn10	2 196	2 222	2 249	2 501	2 278	2 490	2 258	2 293	2 371	2 210
	wn3	2 304	2 393	2 363	2 531	2 339	2 530	2 367	2 324	2 496	2 375
F2 rozp.	c	1 110	982	1 079	1 192	1 249	1 339	999	1 434	1 221	1 373
	wn10	891	984	1 117	1 367	1 179	1 217	1 104	983	1 230	986
	wn3	749	910	1 063	1 095	1 018	1 013	1 024	724	1 037	995
F2 med.	c	1 327	1 370	1 280	1 361	1 326	1 400	1 339	1 327	1 244	1 376
	wn10	1 778	1 682	1 624	1 808	1 660	1 796	1 669	1 800	1 757	1 692
	wn3	1 847	1 799	1 788	1 900	1 793	1 902	1 792	1 855	1 856	1 788
F3 min.	c	1 961	2 269	2 221	2 210	2 085	2 222	2 231	2 193	2 109	2 213
	wn10	2 434	2 464	2 478	2 477	2 313	2 448	2 468	2 449	2 431	2 516
	wn3	2 528	2 523	2 515	2 527	2 432	2 519	2 520	2 526	2 502	2 539
F3 max.	c	3 321	2 794	2 854	2 965	2 812	2 959	2 850	3 063	3 073	2 954
	wn10	3 241	3 190	3 247	3 515	3 260	3 477	3 249	3 280	3 378	3 208
	wn3	3 304	3 370	3 347	3 519	3 336	3 517	3 347	3 302	3 487	3 366
F3 rozp.	c	1 360	525	633	755	727	737	619	870	964	741
	wn10	807	726	769	1 038	947	1 029	781	831	947	692
	wn3	776	847	832	992	904	998	827	776	985	827
F3 med.	c	2 313	2 559	2 571	2 557	2 379	2 523	2 539	2 541	2 431	2 669
	wn10	2 758	2 677	2 694	2 793	2 609	2 743	2 705	2 764	2 733	2 711
	wn3	2 814	2 758	2 781	2 884	2 735	2 867	2 784	2 814	2 833	2 774

Tabulka 5-5: Porovnání hodnot v čisté nahrávce a v nahrávce s bílým šumem (-10 dB a-3 dB) pro první, druhý a třetí formant (v Hz).

5.4 Porovnávání nahrávek s různým šumem, kompenzace šumu

Jelikož se ve forenzní praxi velmi pravděpodobně setkáme s nutností porovnat dvě nahrávky obsahující různý druh a/nebo odstup šumu, je nutné navrhnout metody, pomocí nichž bude možné rozdílnou míru šumu v nahrávkách kompenzovat. Pro další postup uvažujme situaci, kdy máme k dispozici zašuměnou spornou nahrávku

z terénu, u níž je třeba určit mluvčího, a spolupracujícího podezřelého, který je ochoten namluvit srovnávací nahrávku např. ve studiu, kde je šum eliminován.

Srovnáváme tedy čistou kontrolní nahrávku (nahrávka A) a původní nahrávku s prozatím neznámým druhem a odstupem šumu (nahrávka B), jejichž histogramy LTF hodnot se velice pravděpodobně budou lišit, i pokud jde o téhož mluvčího (viz předchozí kapitola). Nabízejí se dva základní možné postupy:

1) Odstranění šumu z nahrávky B.
2) Přidání šumu do nahrávky A.

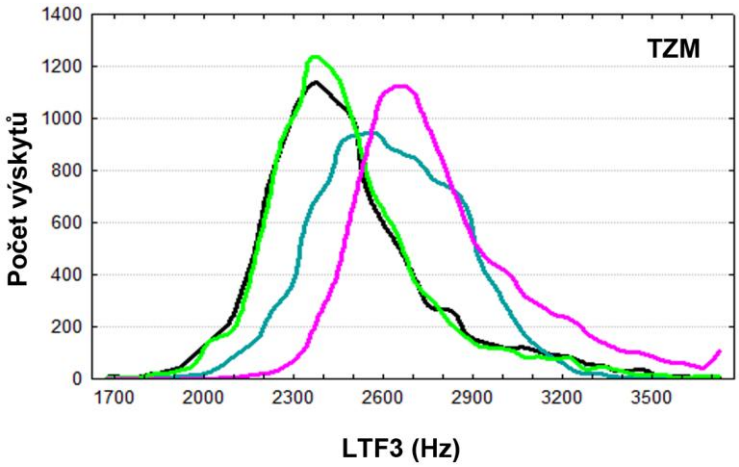
První možnost musíme vyloučit, protože by při odstraňování šumu mohlo dojít i k odstranění některých prvků z řeči mluvčího důležitých pro jeho identifikaci. Abychom mohli šum účinně simulovat, je třeba postupně zodpovědět následující otázky:

- Je možné v nahrávce B identifikovat druh a odstup šumu?
- Jak moc podobný tomuto šumu musí být šum přidaný do nahrávky A, aby histogramy LTF hodnot obou nahrávek vypadaly stejně, v případě, že jde o téhož mluvčího? (To znamená, aby oba šumy zkreslovaly LTF hodnoty stejným způsobem a měrou.)
- Jsme schopni dostatečně podobný šum vytvořit? Jakým způsobem?

Odpověďmi na tyto otázky a návrhem možných postupů se budeme zabývat v následujících kapitolách.

5.4.1 Identifikace druhu a odstupu šumu v nahrávce

Již jsme zmínili, že každý šum deformuje histogram LTF hodnot jiným způsobem (viz obr. 5-11). Rozdíly jsou dokonce patrné i při stejném druhu, ale odlišném odstupu šumu (viz obr. 5-10). Proto je nezbytné šum v nahrávce správně identifikovat, abychom byli schopni jej v čisté nahrávce správně simulovat.

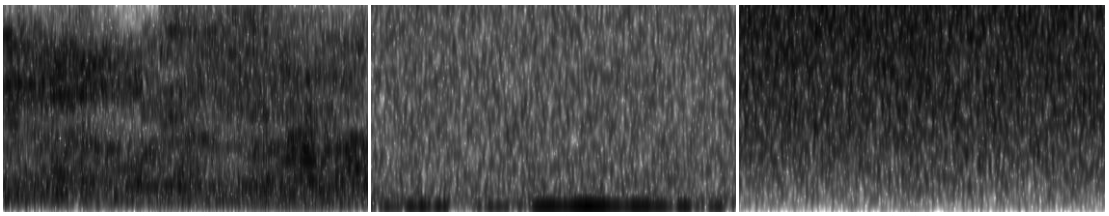


Obrázek 5-11: Vliv různých druhů šumu s odstupem -10 dB na čistou nahrávku (černá): hnědý šum (zelená), kavárenský šum (modrá), bílý šum (růžová). Pro přehlednost jsou znázorněny jen okraje histogramů.

Šumy lze rozdělit do dvou základních skupin:

- terénní – tzn. např. kavárenský šum, hluk na ulici, hučení spotřebiče apod.
- umělé – hnědý, bílý, růžový a další „barevné“ šumy, dále sem můžeme v podstatě zařadit i různé filtry, např. telefonický přenost apod.

Šumy z první skupiny dokáže i laik většinou snadno identifikovat poslechem, ve spektrogramu se nevyznačují ničím zvláštním. Šumy z druhé skupiny může odborník identifikovat jednak poslechem, jednak vizuálně ve spektrogramu. Hnědý šum není ve spektrogramu téměř patrný – výrazně ovlivňuje pouze hodnoty do zhruba 400 Hz (takže je pohledem těžko odlišitelný od F0 a F1), vyšší frekvence pouze slabě překrývá. Bílý šum je naopak snadno rozpoznatelný, poněvadž jeho spektrum přechází od bílé v nejnižších frekvencích do černé ve frekvencích nejvyšších (viz obr. 5-12 s ukázkami samotného šumu a obr. 5-1 na str. 37 s ukázkami nahrávky smíchané se šumem).



kavárenský šum hnědý šum bílý šum

Obrázek 5-12: Spektrogramy kavárenského, hnědého a bílého šumu (frekvenční spektrum 0–5 000 Hz).

Sílu (odstup) šumu v nahrávce je možné přibližně zjistit přes funkci HNR (harmonicity-to-noise ratio). Výsledky nejsou zcela přesné – zdá se, že vypočtený odstup šumu vychází vždy o něco menší, než původní odstup, se kterým byl šum s nahrávkou původně smíchán (pro šum původně -3 dB vychází odstup asi -2 dB, pro -10 dB je to asi -6 dB). Jak ale vyplývá z výsledků této práce, odstup šumu už nehraje při srovnávání nahrávek tak významnou roli. Tvar histogramu bývá většinou obdobný, mění se jen celková frekvenční pozice LTF hodnot (viz kapitola *Vliv různého odstupu a druhu šumu v nahrávce na extrahované LTF hodnoty*, s. 48).

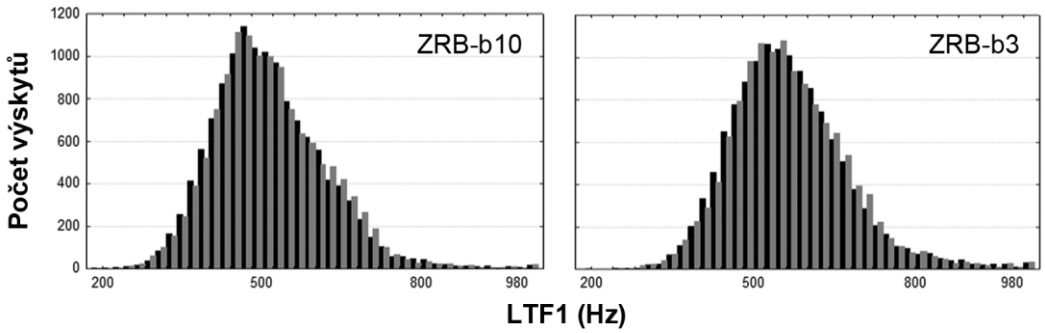
5.4.2 Simulace šumu

Naším cílem je nyní vytvořit šum co nejpodobnější šumu z původní nahrávky. Abychom zjistili, do jaké míry se musí šumy podobat, provedli jsme experiment s posunutým šumem. Ten má odpovědět na otázku, zda hrají roli i drobné detaily šumu (např., vyvedeno do krajnosti, kdy v kavárenském šumu cinkne sklenička), nebo zda jsou rozhodující jen obecné vlastnosti šumu.

Kavárenský šum[3] (-10 dB i -3 dB) jsme upravili tak, že jsme jeho druhou polovinu přesunuli na začátek zvukové stopy. Poté jsme jej smíchali s nahrávkou a vyextrahovali formanty stejným způsobem jako u původních šumů. Tím zůstaly zachovány obecné vlastnosti šumu, ale změnilo se „lícování“ se samotnou nahrávkou – jednotlivé úseky řeči byly tedy nyní ovlivněny jinou částí šumu. Výsledky experimentu jsou pozitivní a příznivé pro další postup práce – histogramy nahrávek s původním

[3] Bílý šum jsme ze všech experimentů vyloučili kvůli výrazné deformaci histogramů LTF hodnot a ztrátě inter-individuální variability LTF hodnot v nahrávkách s bílým šumem. Hnědý šum nebylo třeba testovat, protože už kavárenský šum, který je více „členitý“, obstál v tomto experimentu velice dobře.

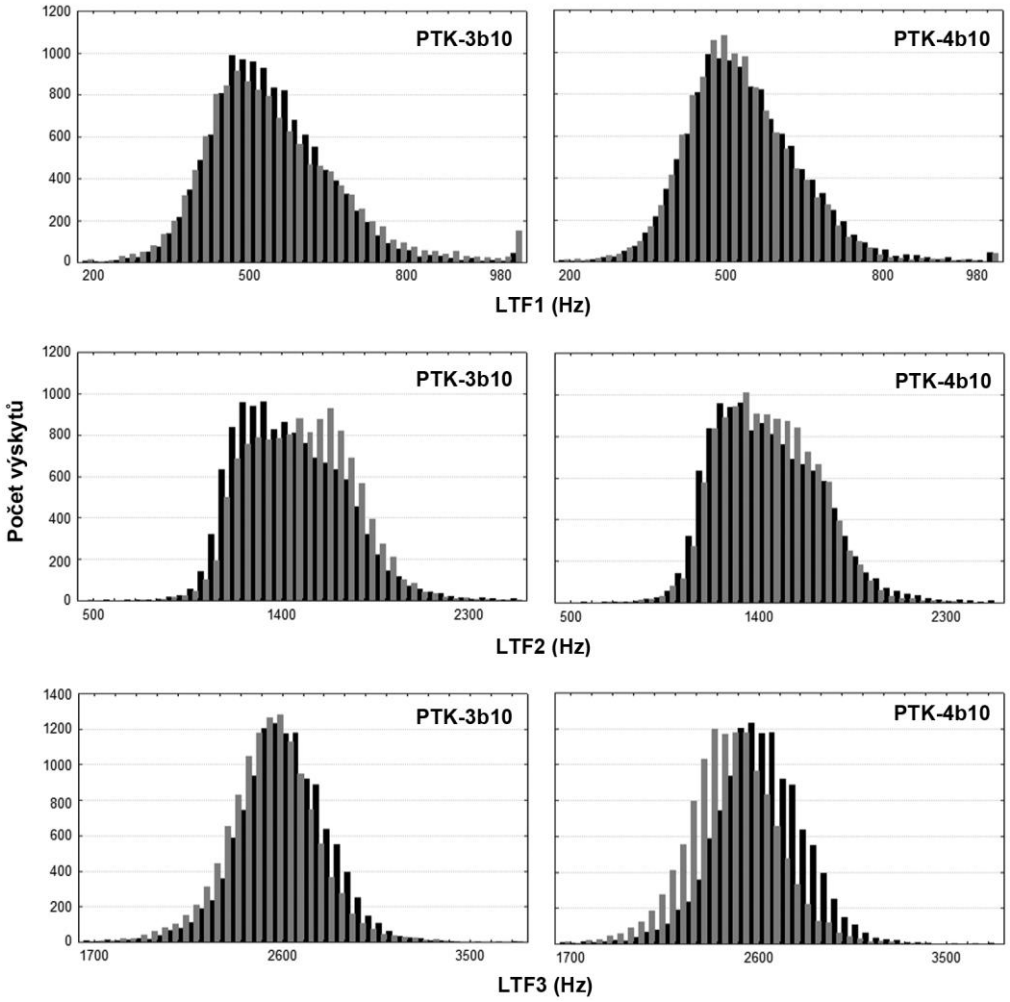
i posunutým šumem jsou prakticky identické, a to pro oba odstupy šumu (obr. 5-13). Znamená to, že drobné detaily v šumu nijak výrazně automatickou extrakci formantů neovlivňují, důležité jsou pouze obecné vlastnosti šumu. Z toho vyplývá, že abychom mohli efektivně porovnávat čistou nahrávku A se zašuměnou nahrávku B, nepotřebujeme získat nutně absolutně stejný šum jako v nahrávce B, ale můžeme se pokusit vytvořit co nejpodobnější šum se stejnými obecnými vlastnostmi.



Obrázek 5-13: Porovnání nahrávky s původním (černá) a posunutým (šedá) kavárenským šumem pro odstup -10 dB (vlevo) a -3 dB (vpravo).

5.4.2.1 Terénní šumy

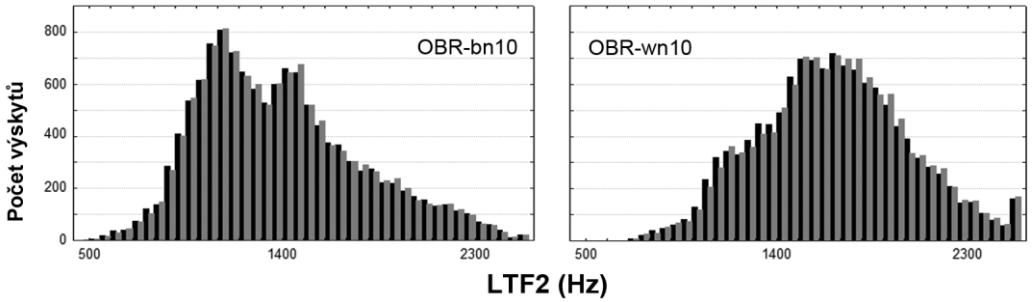
Terénní šumy obsahují velké množství složek, zvuků či hlasů, takže je velice obtížné takový šum vytvořit uměle. Nejsnadnější cestou je stažení volně šiřitelné nahrávky podobného šumu z internetu. Např. kavárenských šumů lze najít přehršel, jen je třeba zajistit, aby byl sluchový dojem co nejpodobnější jako v původním šumu. Vyzkoušeli jsme dva další kavárenské šumy, poslechově podobné původnímu šumu. Výsledky se v obou případech markantně liší. Zatímco histogramy prvního šumu (označení „3b“) se nejvíce shodují s původním šumem v LTF3 (ale dobře použitelný je i LTF1 a s větší tolerancí i LTF2), histogramy druhého šumu (ozněčení „4b“) se směrem k vyšším formantům zhoršují, a to tak, že pro LTF1 jsou histogramy téměř totožné s původními, pro LTF2 se jen lehce liší v oblasti vrcholů, ale pro LTF3 jsou u části mluvčích až nepoužitelné – kromě posunu vrcholu na frekvenční škále se občas mění i celý tvar histogramu. Srovnání obou nově aplikovaných šumů s původním je zachyceno na obr. 5-14. Je tedy zřejmé, že porovnávání dvou nahrávek s různým (byť poslechově podobným) terénním šumem je proveditelné, ale je třeba brát jej s rezervou.



Obrázek 5-14: Srovnání nahrávek se dvěma novými kavárenskými šumy (šedá) s původním kavárenským šumem (černá), odstup šumu -10 dB.

5.4.2.2 Umělé šumy

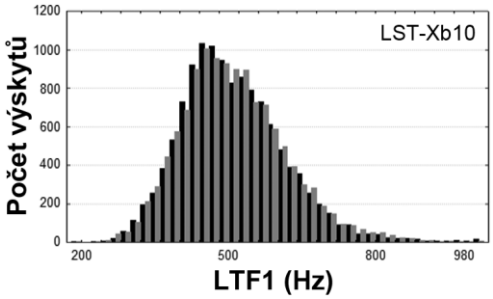
Umělé šumy lze snadno vygenerovat v počítači, důležité je pouze zjistit správný druh šumu (jak vyplývá z předchozího textu, např. vliv hnědého a bílého šumu na nahrávku je naprosto odlišný). Překážkou by neměla být ani různá dodatečná nastavení šumu, jak dokazuje porovnání dvou hnědých šumů vytvořených v programu Cool Edit Pro. Variovali jsme nastavení intenzity šumu (t.j. procento maximální hodnoty vzorku v šumu). Příklad na obr. 5-15 dokládá, že jak pro intenzitu 12, tak 24 jsou histogramy všech mluvčích téměř identické (pro šum -10 dB jsou prakticky totožné, pro -3 dB se objevují zanedbatelné odchylky). Stejně reaguje i bílý šum v zeslabení na intenzitu 2 (odstup -10 dB) – histogramy jsou opět víceméně stejné jako histogramy pro intenzitu šumu 12.



Obr. 5-15: Srovnání LTF2 hodnot v nahrávkách s nově vygenerovaným hnědým (vlevo) a bílým (vpravo) šumem (oba šedá) s nahrávkami s původními šumy stejného druhu (černá), odstup šumu -10 dB.

5.4.3 Extrakce šumu z původní nahrávky

Další možnou alternativou, vhodnou pro oba druhy šumů, je extrakce šumu z původní zašuměné nahrávky. Tato metoda spočívá ve vyseparování tichých pauz (tzn. kdy žádný z mluvčích na nahrávce právě nehovoří ani se nenadechuje; je slyšet jen šum v pozadí). Ideální jsou samozřejmě dlouhé a početné pauzy, takže se tento postup hodí spíše pro delší nahrávky. Minimální celkové trvání vyextrahovaných pauz závisí na druhu šumu – čím variabilnější šum, tím delší úsek šumu je třeba složit. Obecně by vyextrahovaný šum měl trvat alespoň několik sekund, samozřejmě čím čím více, tím lépe.



Obrázek 5-16: Srovnání LTF1 hodnot v nahrávkách s kavárenským šumem s odstupem -10 dB (černá) a s vyextrahovanou pětisekundovou smyčkou téhož šumu (šedá).

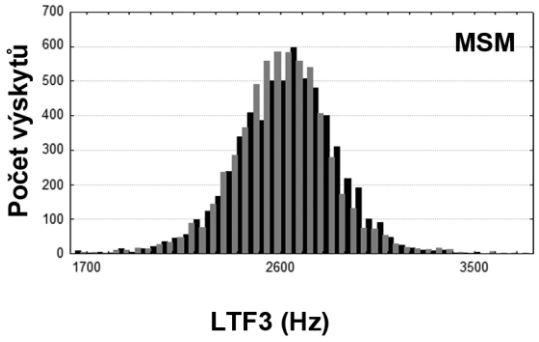
Metodu jsme testovali na kavárenském šumu s odstupem -10 dB. Zhruba pěti-sekundový vyextrahovaný šum (složený z několika pauz) jsme zřetězili, aby pokryl celé trvání 120sekundové nahrávky, a následně vyextrahovali formanty obvyklým

způsobem. Jak je vidět na obr. 5-16, již smyčka o délce 5 sekund je dostačující – histogramy pro LTF1 se šumovou smyčkou se shodují s histogramy nahrávek s původním šumem. Pro LTF2 a LTF3 se histogramy drobně liší v oblasti vrcholů, ale celkový tvar a umístění histogramů na frekvenční škále zůstávají zcela zachovány.

5.4.4 Smíchání zašuměné nahrávky s čistou

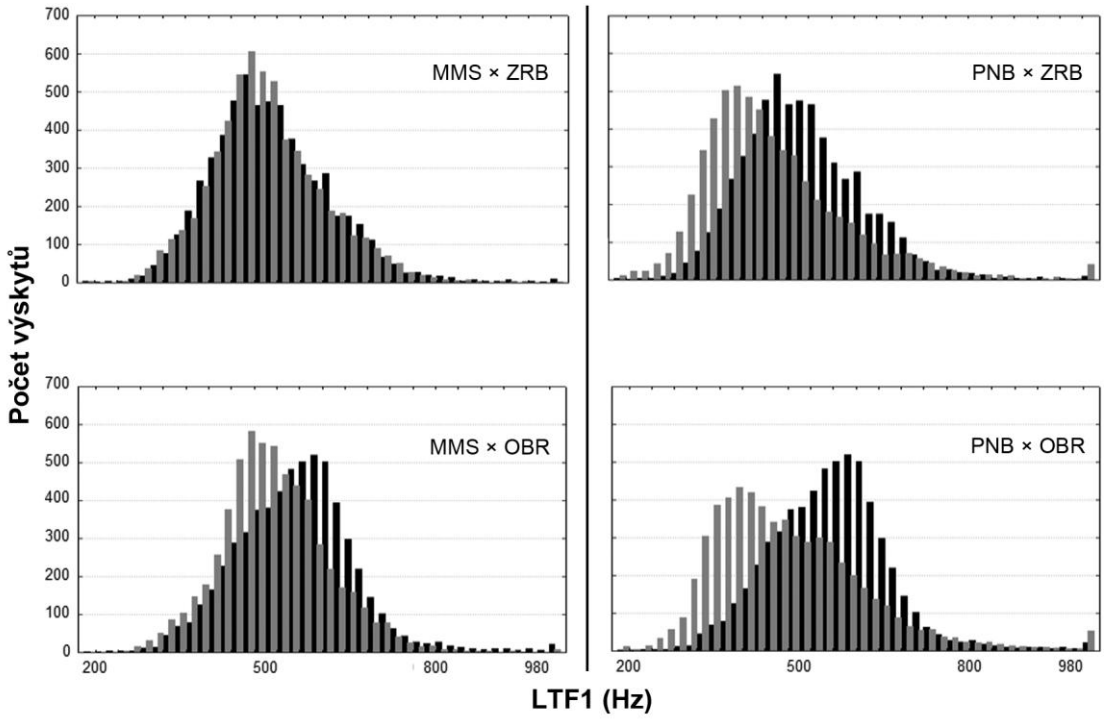
V tomto experimentu jsme vycházeli z předpokladu, že různé nahrávky jednoho mluvčího vykazují stejné histogramy LTF hodnot (pokud jsou dostatečně dlouhé). Disponujeme-li tedy čistou a zašuměnou nahrávkou téhož mluvčího, mělo by být možné je smíchat a získat tak velice podobný histogram jako u zašuměné nahrávky. Ve smíchané nahrávce se budou překrývat oba hlasy, což by ale nemělo vadit, protože ať už se při automatické extrakci formantů vyberou formanty z původní čisté, nebo z původní zašuměné nahrávky, statisticky by celkově měly tvořit stále stejný (nebo alespoň hodně podobný) histogram.

Použili jsme proto 1. minutu z čisté nahrávky (nahrávka A) a 2. minutu ze zašuměné nahrávky (nahrávka B; kavárenský šum, -10 dB). Následně jsme obě nahrávky smíchali (→ nahrávka AB), přičemž nahrávka B měla od „nosné“ nahrávky A odstup -10 dB. Předpokládaný výsledek, tedy že histogram nahrávky AB se bude blížit histogramu nahrávky B, se ukázal u většiny mluvčích (viz příklad na obr. 5-17). Pouze u mluvčího TZM se histogramy všech tří formantů podstatně liší, u dalších tří mluvčích (KCR, PNB a PVC) se výrazněji odlišují histogramy jednoho ze tří formantů. Potud je tedy hypotéza potvrzena.



Obrázek 5-17: Srovnání LTF3 hodnot v nahrávce A (šedá) a B (černá). (Označení je vysvětleno v předchozím odstavci.)

Je ale třeba vzít v potaz i to, že mluvčí na čisté nahrávce A a zašuměné nahrávce B mohou být rozdílní. V tom případě předpokládáme u histogramu AB určité přiblížení k histogramu B, ale měl by zůstat dostatečně odlišný na to, aby bylo možné konstatovat, že jde pravděpodobně o odlišné mluvčí. Tento předpoklad se bohužel nepotvrdil. Testovali jsme kombinaci čistých nahrávek A všech mluvčích se zašuměnou nahrávkou B od mluvčích LST, OBR a ZRB. Mluvčí LST a ZBR vykazují v čisté nahrávce relativně podobné histogramy pro LTF. Co se týče frekvenčních hodnot, patří oba spíše k průměru mezi mluvčími. Naproti tomu histogramy čisté nahrávky mluvčího OBR se od ostatních liší, jednak dvojitými vrcholy u LTF1 a LTF2, jednak frekvenčními hodnotami – hodnoty LTF1 má spíše vyšší, naopak LTF2 a LTF3 spíše nižší než ostatní mluvčí.



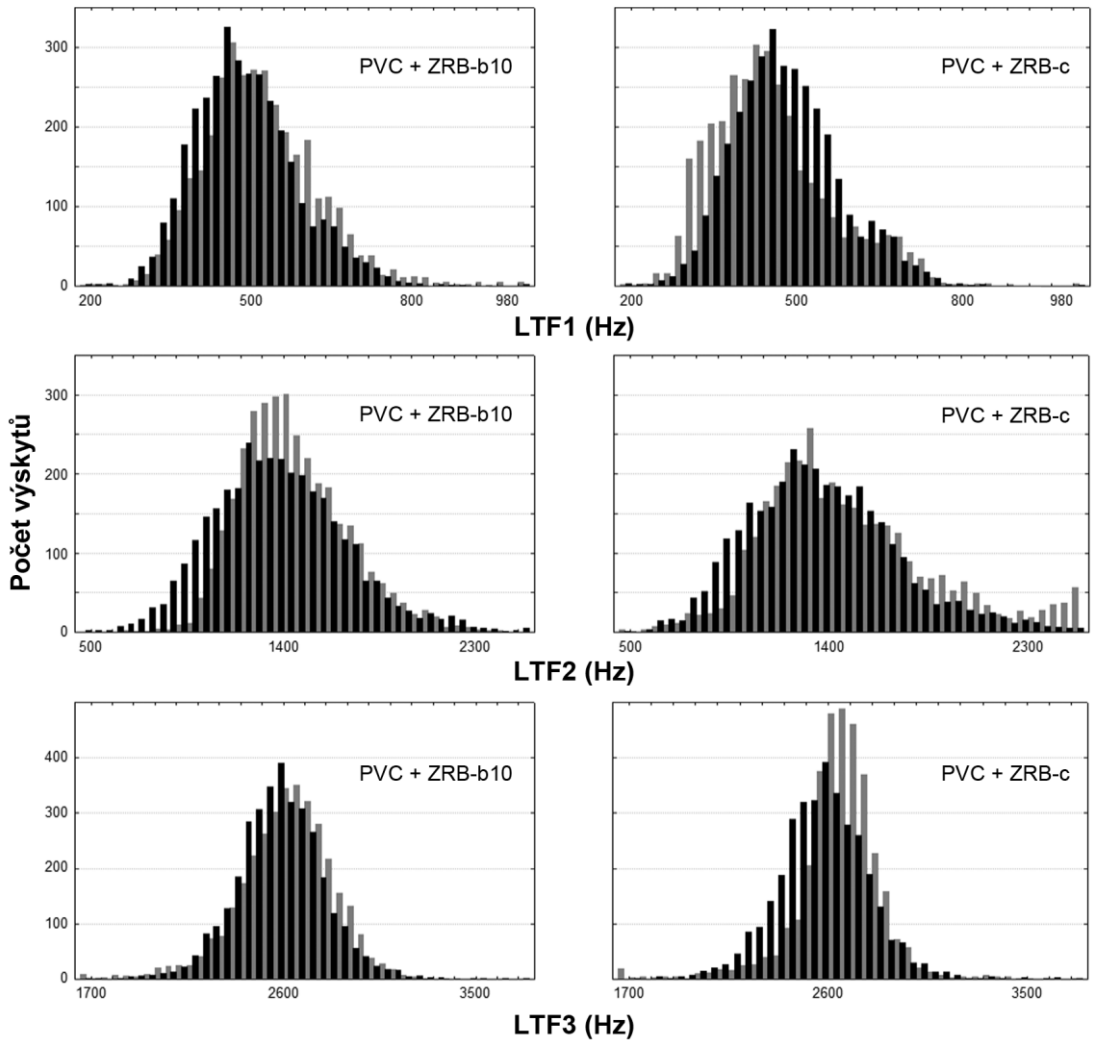
Obrázek 5-18: Srovnání LTF1 hodnot v nahrávkách AB (šedá) mluvčích MMS (vlevo) a PNB (vpravo) a v nahrávkách B (černá) mluvčích ZRB (nahoře) a OBR (dole). Zatímco mluvčí PNB se znatelně liší od od obou porovnávaných mluvčích, histogram mluvčího MMS je téměř totožný s mluvčím ZRB, což znamená, že použitá metoda nefunguje. (Označení nahrávek je vysvětleno na předchozí straně.)

Ukázalo se, že přimíchání zašuměné nahrávky B jiného mluvčího do čisté nahrávky A ovlivňuje tuto natolik, že v mnohých případech není její histogram rozeznatelný od histogramu samotné nahrávky B. To znamená, že původní čistá nahrávka po

překrytí nahrávkou jiného mluvčího ztrácí své původní vlastnosti (u některých mluvčích částečně, u jiných zcela) a přebírá je od přidaného šumu. Po smíchání s mluvčími LST a ZRB byly pro velkou většinu mluvčích histogramy LTF1 a LTF3 velmi podobné histogramům samotných nahrávek B mluvčích LST a ZRB (tedy nahrávek jiných mluvčích). Odlišovali se pokaždé pouze asi 2–3 mluvčí. Alespoň v histogramech LTF2 můžeme vidět větší odlišnosti. Je pozoruhodné, že někteří mluvčí se s „cizími" nahrávkami B shodovali více než sami mluvčí nahrávek B. Při porovnávání s mluvčím OBR můžeme oproti tomu vidět nejmarkantnější rozdíly v LTF1 (viz obr. 5-18), což je zapříčiněno právě tím, že LTF1 mluvčího OBR leží ve vyšších frekvencích než u jiných mluvčích. U LTF3, a především LTF2 se rozdíly opět stírají a z hstogramů nelze jednoznačně poznat, že jde o dva různé mluvčí.

Pro zajímavost jsme provedli ještě jeden obdobný pokus, tentokrát jsme ale pro smíchání s čistou nahrávkou všech mluvčích použili jen čistou nahrávku mluvčího ZRB (opět s odstupem -10 dB). Cílem bylo zjistit, do jaké míry ovlivňuje výsledky předchozího experimentu šum v přidané nahrávce a do jaké míry samotný hlas v přidané nahrávce. Přestože histogramy některých mluvčích zůstávají dost podobné samotné čisté nahrávce mluvčího ZRB (především pro LTF2), mnohem více histogramů se teď odlišuje výrazněji. I přesto lze ale nalézt histogramy, které jsou mluvčímu ZRB podobnější než mluvčí ZRB sám sobě. Opět zde ale funguje alespoň frekvenční umístění vrcholu histogramu. Mluvčí ZRB má nejvyšší LTF3 ze všech mluvčích, což je vidět i v porovnání těchto histogramů – žádný „smíchaný" histogram nemá vrchol frekvenčně výše než sám mluvčí ZRB.

Na obrázku 5-19 je vidět, že na histogram mluvčího PVC nemá příliš vliv, zda je v šumové nahrávce B přítomen pouze hlas mluvčího ZRB, nebo i šum – mění se pouze tvar srovnávací nahrávky mluvčího ZRB. Je tedy patrné, že samotný šum, jenž má ve smíchané nahrávce odstup -20 dB (-10 dB v nahrávce B, dalších -10 dB po smíchání s nahrávkou A), nahrávku ovlivňuje jen minimálně, největší podíl na změně LTF hodnot má „šumový" hlas z nahrávky B.



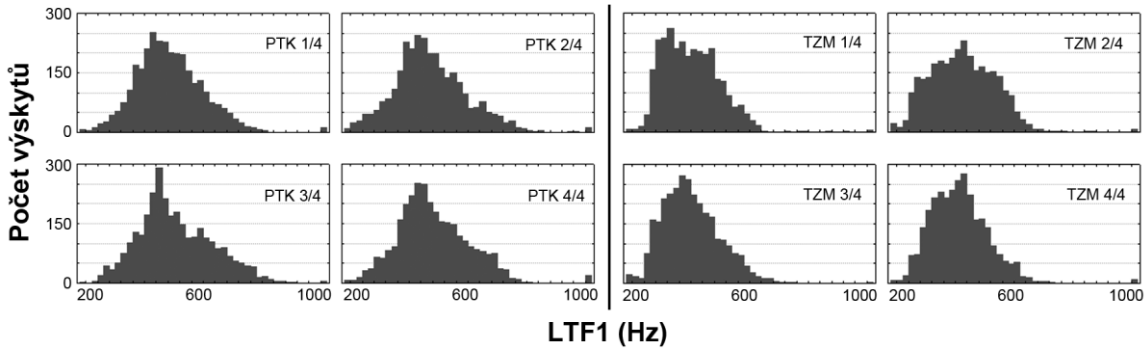
Obrázek 5-19: Porovnání histogramů čisté nahrávky mluvčího PVC (černá) zašuměné nahrávkou mluvčího ZRB (šedá) s kavárenským šumem (vlevo) a zašuměné čistou nahrávkou mluvčího ZRB (vpravo), odstup šumu -10 dB.

5.5 Minimální trvání vokalického proudu

Na několika místech textu operujeme s pojmem „dostatečně dlouhá nahrávka“. Je to taková nahrávka, která svým trváním, resp. počtem naměřených hodnot přesahuje určitou nejnižší hranici. Všechny nahrávky jednoho mluvčího, které tuto hranici přesahují, by měly za totožných podmínek vykazovat shodná rozdělení LTF hodnot, maximálně s drobnými odchylkami. Kratší nahrávky s nedostatečným počtem hodnot mohou zahrnovat jen část spektra LTF hodnot, a nemají tudíž patřičnou vypovídající hodnotu. (V případě nouze, kdy není možné zajistit delší nahrávku, je třeba brát takovéto výsledky s rezervou.) Ve forenzní praxi je často bohužel nutné pracovat

s nedostatečně dlouhými nahrávkami, je tedy třeba zjistit, jak se kratší nahrávky chovají a do jaké míry se od sebe mohou lišit nahrávky téhož mluvčího určitého trvání.

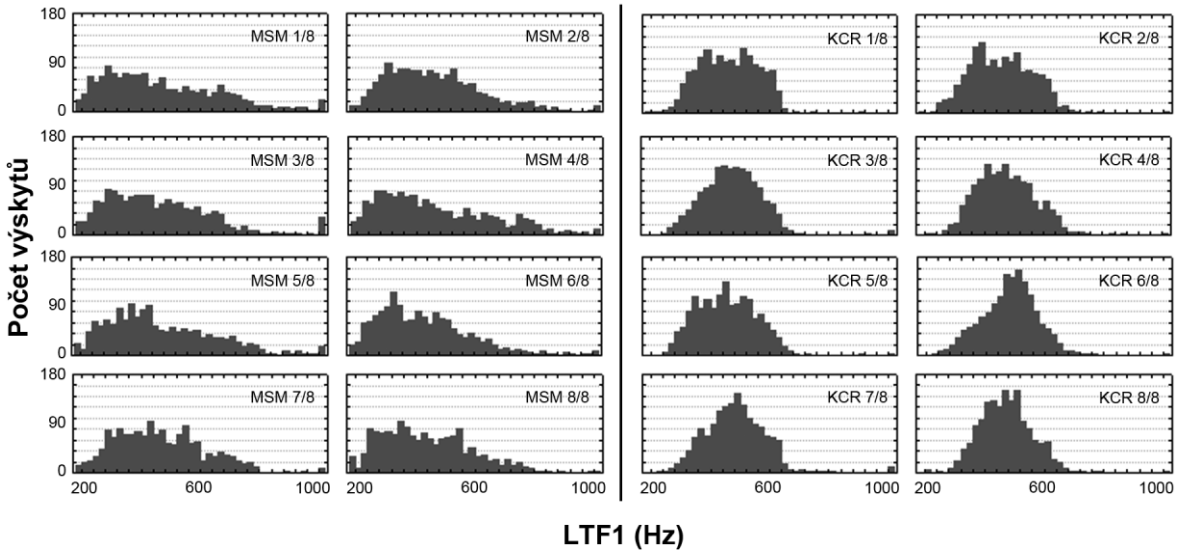
Porovnání polovin nahrávek (tedy 2× cca. 60 s) jsme provedli již v rámci zjišťování intra-individuální variability mluvčích (viz kapitola *Intra-individuální variabilita*, s. 38). Příklady na obr. 5-20 znázorňují srovnání čtvrtin nahrávek (4× cca. 30 s). Zatímco pro mluvčího PTK vycházejí všechny čtyři histogramy hodnot LTF1 velice podobně, u mluvčího TZM se znatelně liší druhá čtvrtina (změna špičatosti a sešikmení) a o něco méně i první (změna sešikmení). Přesto ale u všech mluvčích zůstává zachováno frekvenční umístění histogramu a v podstatě se nemění ani okraje histogramů – ke změnám tvaru dochází (pokud vůbec) především v oblasti vrcholu histogramu. U žádného mluvčího ale nejde o tak výrazné změny, že by zcela znemožnily identifikaci, např. nikde nedochází k úplnému otočení sešikmení. Většinou se navíc od ostatních odlišuje jen jedna ze čtvrtin, takže pravděpodobnost, že 30sekundové nahrávky pocházející od jednoho mluvčího budou mít stejné rozložení LTF hodnot, je relativně vysoká. Pro F2 a F3 je shoda mezi čtvrtinami nahrávek ještě lepší.



Obrázek 5-20: Vzájemné srovnání hodnot LTF1 v čistých nahrávkách rozdělených na čtyři části. Zatímco u mluvčího PTK (vlevo) si jsou všechny histogramy podobné, u mluvčího TZM (vpravo) se výrazně odlišuje druhá a částečně i první čtvrtina.

Osminové části nahrávek (cca. 15 s) bohužel vykazují již větší množství vzájemných odlišností. Pro LTF1 zůstává (s drobnými odchylkami) zachován víceméně pouze nejstabilnější rys histogramů, totiž jejich rozsah na frekvenční škále. Jinak se mění špičatost a počet vrcholů histogramu (např. mluvčí LST), dochází k posunu vrcholu histogramu, a to např. u mluvčího KCR dokonce do takové míry, že se zcela

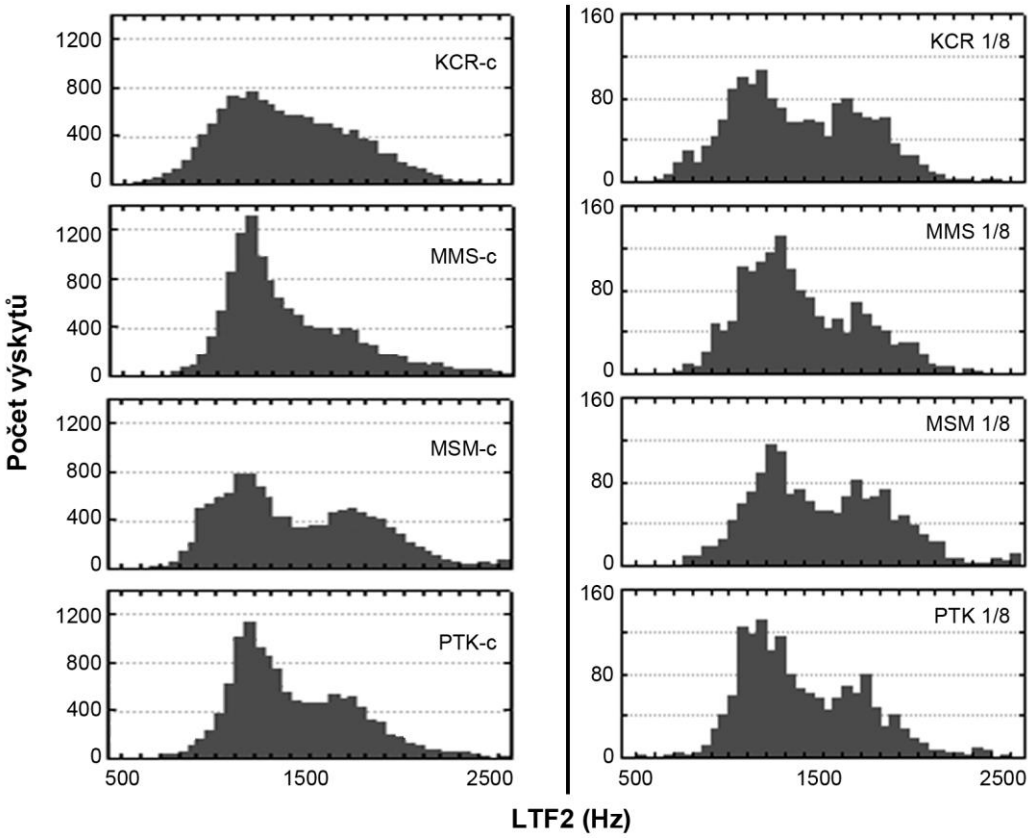
otáčí směr sešikmení grafu. Mezi osminami lze vždy nalézt některé, které si tvarem odpovídají, ale počet odlišných oproti čtvrtinám vzrostl. (viz obr. 5-21). Identifikace mluvčích už je zde značně ztížena, neboť můžeme bezpečně porovnávat pouze frekvenční rozložení LTF hodnot, ostatní faktory mohou být znatelně odlišné i u jednoho mluvčího.



Obrázek 5-21: Vzájemné srovnání hodnot LTF1 v čistých nahrávkách rozdělených na osm částí. Rozdělení hodnot u mluvčího MSM (vlevo) zůstává relativně zachováno, kdežto některé z osmin mluvčího KCR (vpravo), např. první a šestá, se již výrazně liší.

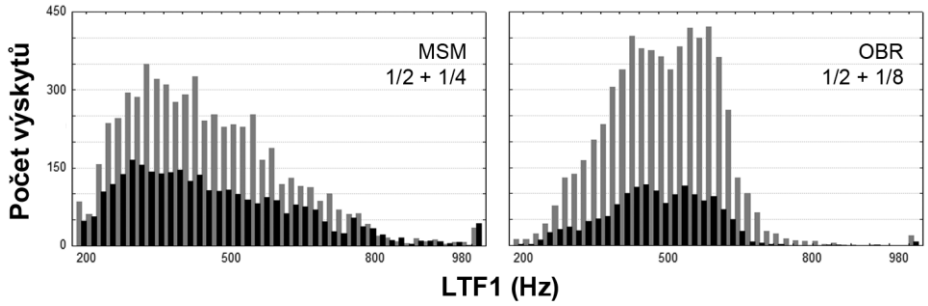
Míra vizuální shody histogramu klesla u osmin nahrávek i pro LTF2. Ačkoliv pro některé mluvčí (např. MMS) zůstává všech osm histogramů dostatečně shodných (liší se jen ve špičatosti grafu), u jiných (např. PNB) můžeme pozorovat i odlišnosti ve sklonu či počtu vrcholů grafu. Na obr. 5-22 si můžeme všimnout, jak se se zkracujícím se trváním nahrávky (= se snižováním počtu extrahovaných hodnot) vytrácí inter-individuální variabilita hodnot u mluvčích – zatímco v plné délce nahrávky (120 s, asi 12 000 extrahovaných hodnot) jsou histogramy mluvčích viditelně odlišné, při osminovém trvání nahrávky (15 s, asi 1 500 extrahovaných hodnot) se některé histogramy jednotlivých mluvčích začínají navzájem znatelně podobat.

Histogramy LTF3 se liší téměř výhradně pouze ve špičatosti histogramu – frekvenční rozsah a poloha vrcholu zůstávají zachovány. Obecně se LTF3 zdá proti zkracování trvání nejodolnější a lze předpokládat, že by u většiny mluvčích v případě nutnosti snesl ještě větší zkrácení.



Obrázek 5-22: Ukázka částečné ztráty inter-individuální variablity u hodnot LTF2 vlivem nedostatečného trvání nahrávky. (Ukázky vlevo mají trvání asi 120 s, resp. asi 12 000 extrahovaných hodnot, ukázky vpravo pak pouze 15 s a 1 500 hodnot).

Teoreticky je možné v případě potřeby porovnávat i nahrávky rozdílného trvání. V takovém případě je třeba mít na paměti, že histogram kratší nahrávky by neměl nikde výrazně „vyčnívat“ z histogramu delší nahrávky (viz obr. 5-23). Důvod je logický – pokud trvá kratší nahrávka např. čtvrtinu času delší nahrávky, měla by i v grafu zabírat čtvrtinu celkových hodnot, a sice na stejné (případně o něco málo menší) frekvenční škále a v nižších četnostech než celková nahrávka. (Samozřejmě je třeba počítat s tím, že drobné odchylky v histogramech se vyskytují i mezi delšími nahrávkami se stejným trváním, takže případné malé přesahy nad rámec histogramu celkové nahrávky lze tolerovat.) Pro experiment jsme vybrali vždy druhou polovinu nahrávky (asi 60 s) a k ní čtvrtinu, resp. osminu (asi 30 a 15 s) z první poloviny, abychom neporovnávali naprosto stejné části.



Obrázek 5-23: Porovnání dvou nahrávek téhož mluvčího s různým trváním (vlevo 60 a 30 s, vpravo 60 a 15 s).

Protože zmíněná trvání se vztahují k vokalickým proudům, na závěr se podíváme na to, jak dlouhá musí být samotná nahrávka řeči (s již odstraněnými pauzami) před selekcí sonorních hlásek (viz tabulka 5-6). V prvním sloupci je uvedeno celkové trvání nahrávky, z níž byl extrahován vokalický proud (s trváním ve druhém sloupci). Pokud obě čísla vydělíme, získáme koeficient, který udává, kolik nevyužitého materiálu připadá na jednu sonorní jednotku. Minimální trvání řeči tedy získáme tak, že vynásobíme tento koeficient minimálním trváním vokalického proudu.

	původní řeč	původní vokal. proud	koeficient	minimální trvání řeči
KCR	288 s	124 s	2,32	**34,84 s**
LST	270 s	123 s	2,20	**32,93 s**
MMS	257 s	121 s	2,12	**31,86 s**
MSM	282 s	121 s	2,33	**34,96 s**
OBR	291 s	124 s	2,35	**35,20 s**
PNB	296 s	122 s	2,43	**36,39 s**
PTK	280 s	123 s	2,28	**34,15 s**
PVC	144 s	69 s	2,09	**31,30 s**
TZM	282 s	120 s	2,35	**35,25 s**
ZRB	302 s	129 s	2,34	**35,12 s**

Tabulka 5-6: Výpočet minimálního nutného trvání nahrávky řeči (bez pauz) ze zjištěného minimálního trvání vokalického proudu (15 s).

Zde je nasnadě srovnání se studií Anji Moos (2008), která odhadla minimální možné trvání nahrávky na 6–8 s v závislosti na formantu. Oproti hodnotám odhadnutým v této práci je to zhruba poloviční trvání. Je ovšem třeba podotknout, že Moos pracovala pouze s průměrnou LTF hodnotou a směrodatnou odchylkou, jež je snad možno vypočítat i z menšího množství dat, zatímco LTF distribuce je na krácení trvání náchylná mnohem více, neboť změny v tvaru rozložení hodnot se začnou projevovat už při menším zkrácení.

6 Závěr

Tato diplomová práce si kladla za cíl zjistit potenciál využití metody dlouhodobé formantové distribuce (LTF) ve forenzně-fonetické praxi, především pak u nahrávek se zhoršenou akustickou kvalitou – se šumem v pozadí. Prověřili jsme, do jaké míry je LTF distribuce intra- a inter-individuální a zda je tedy vůbec pro identifikaci mluvčích vhodná. Poté jsme prozkoumali změny, k nimž dochází v extrakci formantů vlivem přítomného šumu, a navrhli jsme metody, pomocí nichž je možné kompenzovat rozdílnou sílu či druh šumu ve dvou porovnávaných nahrávkách. Nakonec jsme se zaměřili na zjištění minimálního trvání nahrávky, které již může poskytnout plnohodnotné výsledky. Na tomto místě bychom rádi shrnuli výsledná zjištění.

Šum v nahrávce nepůsobí pouze na extrahované LTF hodnoty, ale ztěžuje i samotnou přípravu materiálu. V závislosti na druhu šumu a jeho odstupu od nahrávky se spektrogram stává hůře čitelný až nečitelný. Z použitých šumů nahrávku nejméně (téměř vůbec) ovlivňuje hnědý šum. U kavárenského šumu již pozorujeme patrné zhoršení kvality nahrávky a nahrávka s bílým šumem v pozadí je prakticky nezpracovatelná. Vliv šumu je obecně nepřímo úměrný odstupu nahrávky od šumu (SNR).

U LTF distribucí jsme zjistili vysokou vizuální intra-individuální stabilitu – srovnání první a druhé půlky nahrávky u jednotlivých mluvčích ukazuje, že oba histogramy rozdělení LTF hodnot jsou prakticky totožné (pro všechny tři formanty). Dokázali jsme tím, že pokud máme k dispozici dostatečně dlouhé nahrávky od jednoho mluvčího, histogramy LTF hodnot si budou navzájem velice podobné. Pokusili jsme se ověřit toto zjištění i statisticky pomocí Kolmogorovova-Smirnovova testu. Bohužel musíme konstatovat, že test je pro takto obsáhlé soubory hodnot příliš citlivý a toleruje pouze nepatrné odchylky mezi porovnávanými datasety. Test byl úspěšný

pouze v případě, kdy jsme rozpůlili data z nahrávky nikoliv časově, ale náhodně, respektive na sudé a liché vzorky. Takovýto způsob rozdělení dat už obstál i v tomto senzitivním testu. To alespoň dokazuje, že automatický extrakční algoritmus funguje správně a extrahuje skutečné formantové trajektorie.

I zjištěnou inter-individuální variabilitu lze považovat za dostatečnou. Představili jsme několik metod zobrazení a srovnávání LTF hodnot, jež se liší mírou výpovědní hodnoty (a nepřímo úměrně tomu úsporností zobrazení). Tabulkové zobrazení percentilových či středních hodnot je vhodné pro další statistickou analýzu. Pro hrubé srovnání velkého množství mluvčích a vytvoření populačních korpusů se hodí bodové grafy vyjadřující závislosti jednotlivých formantů. Ty ovšem zobrazují pouze střední LTF hodnoty (medián či průměr) a nijak nereflektují podrobnosti tvaru rozložení hodnot. Krabicový graf umožňuje už detailnější srovnání tvaru rozdělení (zobrazuje střední hodnotu, špičatost a sešikmení). Nejlepší porovnání LTF hodnot nabízí histogramy, které jsou ovšem prostorově náročné a při větším počtu mluvčích i nepřehledné. Všechny použité metody ale odhalují významné rozdíly v LTF distribucích jednotlivých mluvčích. Nelze sice předpokládat, že by každý mluvčí měl unikátní histogram, výsledky lze ovšem použít pro rychlou kategorizaci mluvčích a rozhodnutí, zda má smysl pokračovat v další analýze pomocí jiných metod.

Při srovnávání čistých nahrávek s nahrávkami se šumem se ukázalo, že přítomnost šumu v nahrávce má vliv na extrahované LTF hodnoty. Obecně lze říci, že vliv hnědého šumu je minimální, vliv kavárenského šumu znatelný a z nahrávek s bílým šumem se vytrácí inter-individuální variabilita. Čím hlasitější je šum vůči nahrávce, tím více hodnoty ovlivňuje. Bohužel se nám nepodařilo vysledovat žádné systematické změny, každý mluvčí na přítomnost každého šumu reaguje různě (pokud vůbec) – od změny počtu vrcholů přes změny špičatosti, sešikmení a polohy těžiště hodnot až k posunu celého histogramu na frekvenční škále. Jako nejstabilnější rys lze označit právě frekvenční rozsah histogramu – většina změn se odehrává jen kolem oblasti těžiště, okraje histogramu, a tedy i jeho frekvenční rozsah, se mění jen výjimečně. Jediným společným rysem vlivu šumu na LTF hodnoty je obecně posun do vyšších frekvencí – to může být ovšem způsobeno i druhem šumu. Jako nejstabilnější formant se ukázal LTF3, který na šum reaguje nejméně (s výjimkou bílého šumu).

Aby bylo možné srovnávat nahrávky s různým typem šumu (typicky zašuměná sporná nahrávka a čistá srovnávací), představili jsme metody možné kompenzace šumu. Pozitivním zjištěním je, že nahrávku neovlivňují detaily šumu, ale jeho celkové

vlastnosti (o čemž svědčí shodné histogramy nahrávek s posunutým šumem). Jestliže jsme tedy schopni vytvořit dostatečně podobný šum, můžeme jej smíchat s čistou nahrávkou a získáme tak dva porovnatelné vzorky. Slibnou metodou je simulace šumu, kdy ze zašuměné nahrávky zjistíme typ a odstup šumu a v editačním audio programu vytvoříme šum se stejným nastavením (v případě umělých šumů) či obdobný šum nahrajeme, popř. stáhneme z internetu (v případě terénních šumů). Pro umělé šumy je tato metoda spolehlivá, pro terénní šumy je třeba počítat s určitými nepřesnostmi způsobenými variabilitou šumu. Druhou metodou, vhodnou spíše pro delší nahrávky s častějšími pauzami, je extrakce šumu z tichých pauz sporné nahrávky. V závislosti na variabilitě šumu je dostačujících již několik sekund extrahovaného šumu. Ten se nahraje do smyčky a následně se použije na čistou nahrávku.

Posledním řešeným tématem bylo minimální požadované trvání nahrávky pro extrakci spolehlivých LTF distribucí. Ze srovnání polovin, čtvrtin a osmin nahrávek vyplývá, že již při 30sekundové nahrávce se u některých mluvčích začínají projevovat změny v histogramu. V 15sekundové nahrávce jsou již změny častější a výraznější a proto obecně nelze příliš doporučit používání nahrávek ještě kratších – to při 15 s vokalického proudu znamená minimálně zhruba 30–40 s řeči (bez pauz). Tato hodnota je ovšem pro každého mluvčího individuální, není tedy možné určit přesnou minimální hranici. Nejvhodnější se opět zdá být formant LTF3, na němž se zkracování trvání projevuje nejméně. U tohoto formantu bychom si pravděpodobně mohli dovolit zkrátit trvání o další polovinu, tedy na cca. 8 s. Nejstabilnějším rysem je i zde frekvenční rozpětí rozložení LTF hodnot.

Téma využití metody LTF ve forenzní praxi samozřejmě není vyčerpáno a i tato práce nastínila nové otázky, na něž by budoucí výzkum mohl odpovědět. Především je třeba vytvořit dostatečně početný populační korpus LTF hodnot (a tvarů LTF distribucí) jednotlivých mluvčích, aby bylo možné určit obvyklost zjištěných hodnot v populaci. Bylo by vhodné otestovat navržené metody kompenzace šumu na skutečných případech z praxe. Ačkoliv to nyní vypadá nepravděpodobně, ve způsobu, jakým šum ovlivňuje LTF hodnoty jednotlivých mluvčích, může být hlubší skrytý systém. K prozkoumání by ale bylo třeba větší základny mluvčích. Detailnější změření si zaslouží i nutné minimální trvání vokalického proudu.

Reference

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic speaker verification using formant features and gaussian mixture models. *Proceedings of Interspeech 2008*, 1505–1508.

Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer (Version 5.3.55) [Software]. Retrieved 2nd September , 2013, from http://www.praat.org

Braun, A. (1995). Fundamental frequency – how speaker-specific is it? In Braun, A. & Köster, J.-P. (Eds.). *Studies in Forensic Phonetics: Beiträge zur Phonetik und Linguistik, 64,* 9–23.

Braun, A. (1996). Age estimation by different listener groups. *The International Journal of Speech, Language and Law, 3,* 65–73.

Bricker, P. D. & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40:* 1441–1450.

Broderick, P. K., Paul, J. E., & Rennick, R. J. (1975). Semi-automatic speaker identification system. Carnahan conference on crime countermeasures, Lexington, University of Kentucky. May 7–9.

Bull, R. & Clifford, B. (1984). Earwitness voice recognition accuracy. In G. Wells and E. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 92–123). New York: Cambridge University Press.

Butcher, A. (2002). Forensic Phonetics: Issues in speaker identification evidence. *Proceedings of the Inaugural International Conference of the Institute of Forensic Studies: "Forensic Evidence: Proof and Presentation"*. Prato, Italy. July 3–5.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE, 85,* 1437–1462.

Český statistický úřad (2013). Soudnictví, kriminalita, nehody. *Statistická ročenka České republiky 2013* [Online]. Retrieved from http://www.czso.cz/csu/2013edicniplan.nsf/kapitola/0001-13-r_2013-2700

Doherty, E. T. & Hollien, H. (1978). Multiple-factor speaker identification of normal and distorted speech. *Journal of Phonetics, 6,* 1–8.

Doty, N. D. (1998). The influence of nationality on the accuracy of face and voice recognition. *American Journal of Psychology, 111,* 191–214.

Ellis, S. (1994). The Yorkshire Ripper enquiry: Part 1. *Forensic Linguistics, 1,* 197–206.

Eriksson, A. (2005). Tutorial on forensic speech science. Part I: Forensic phonetics. In Interspeech 2005 - Eurospeech 2005. *Proceedings of the 9th European conference on speech communication and technology*. Lisbon, Portugal. September 4–8.

Foulkes, P. & French, J.P. (2012). Forensic speaker comparison: A linguistic-acoustic perspective. In Tiersma, P. & Solan, L. (Eds.), *Oxford handbook of language and law* (pp. 557–572). Oxford: Oxford University Press.

French, J. P., Harrison, P, & Lewis, J. W. (2006). R -v- John Samuel Humble: The Yorkshire Ripper hoaxer trial. *International Journal of Speech, Language and the Law, 13,* 255–273.

Glenn, J. W. & Kleiner, N. (1968). Speaker identification based on nasal phonation. *Journal of the Acoustical Society of America, 43,* 368–372.

Goldstein, U. G. (1976). Speaker-identifying features based on format tracks. *Journal of the Acoustical Society of America, 59,* 176–182.

Goldstein, A. G., Knight, P. et al. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society, 17,* 217–220.

Greisbach, R. (1999). Estimation of speaker height from formant frequencies. *Forensic Linguistics, 6(2),* 265–277.

Grey, G. & Kopp, G. A. (1944). Voiceprint identification. *Bell Telephone Laboratories Report,* 1–14.

Hirson, A., French, J. P., & Howard, D. (1995). Speech fundamental frequency over the telephone and face-to-face: Some implications for forensic phonetics. In Lewis, J. W. (Ed.), *Studies in general and English phonetics* (pp. 230–240). London: Routledge.

Hollien, H. (1990). *The accounts of crime: The new science of forensic phonetics*. New York: Plenum Press.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voice under normal, stress, and disguised speaking condition. *Journal of Phonetics, 10,* 139–148.

Hollien, H. & Schwartz, R. (2000). Aural-perceptual speaker identification: Problems with noncontemporary samples. *Forensic Linguistics, 7,* 199–211.

Hollien, H. & Schwartz, R. (2001). Speaker identification utilizing noncontemporary speech. *Journal of Forensic Sciences, 46,* 63–67.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. , & Nolan, F. (2007). F0 statistics for 100 young male speakers of standard Southern British English. *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1809–1812). Saarbrücken, Germany. August 6–10.

Jazyková poradna ÚJČ AV ČR, v.v.i. (2015). *Internetová jazyková příručka* [Online]. Retrieved from http://prirucka.ujc.cas.cz

Jessen, M. (2010). The forensic phonetician: Forensic speaker identification by experts. In Coulthard, M. & Johnson, A. (Eds.), *Routledge handbook of forensic linguistics* (pp. 378–394). London: Routledge.

Jessen, M. & Becker, T. (2010). Long-term formant distribution as a forensic-phoneitc feature. *2nd Pan-American/Iberian Meeting on Acoustics*. Lecture conducted from ASA, Cancún.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *Journal of Speech, Language and the Law, 12 (2)*, 174–213.

Kersta, L. G. (1962). Voiceprint identification. *Nature, 196*, 1253–1257.

Künzel, H. J. (1987). Sprechererkennung: Grundzüge forensischer Sprachverarbeitung. Heildeberg: Kriminalistik Verlag.

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight?. *Phonetica, 46,* 117–125.

Künzel, H. J. (1990). *Phonetische Untersuchungen zur Sprechererkennung durch linguistisch naive Personen*. Stuttgart: Steiner.

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *International Journal of Speech, Language and the Law, 7(2),* 149–179.

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics, 8,* 80–99.

Künzel, H. (2002). Rejoinder to Francis Nolan's 'The "telephone effect" on formants: A response'. *Forensic Linguistics, 9 (1),* 83–86.

Machač, P. & Skarnitzl, R. (2009). *Fonetická segmentace*. Praha: Epocha.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, *13*, 89–126.

McDougall, K., Nolan, F., Harrison, P., & Kirchhübel, C. (2012). Characterising speakers using formant frequency information: A comparison of vowel formant measurements and long-term formant analysis. *Proceedings of IAFPA 2012*.

McGehee, F. (1937). The reliability of the identification of the human voice. *J. Gen. Psychology, 17*, 249–271.

McGehee, F. (1944). An experimental study of voice recognition. *J. Gen. Psychology, 31*, 53–65.

Meuwly, D. (2003). Le mythe de « L'empreinte vocale » (I). *Revue internationale de criminologie et de police technique et scientifique, 56(2)*, 219–236.

Moos, A. (2008). *Forensische Sprechererkennung mit der Messmethode LTF (long-term formant distribution)* (Unpublished master thesis). Universität des Saarlandes, Saarbrücken, Germany.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician,101,* 7–24.

Morisson, G. S. (2011). Forensic voice comparison and the paradigm shift in forensic science. *Expert Evidence Conference*. Lecture conducted from National Judical College of Australia and the ANU College of Law, Canberra.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1999). Speaker identification and forensic phonetics. In Hardcastle, W. J. & Laver, J. (Eds.), *Handbook of phonetic sciences*. Oxford: Blackwell.

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. *Proceedings of the conference Law and Language: Prospect and Retrospect*. Levi, Finland. December 12–15.

Nolan, F. (2002). The 'telephone effect' on formants: A response. *Forensic Linguistics, 9,* 74–82.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identication. *Speech, Language and the Law, 12 (2),* 143–173.

Orchard, T. L. & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9(3),* 249–260.

Phonetisches Laboratorium (UZH) (2014). *IAFPA 2014, Programme*. Retrieved from http://www.pholab.uzh.ch/iafpa2014.html

Pollák, P., Volín, J., & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. *Proceedings of the XIIth international conference Speech and computer – SPECOM 2007* (pp. 537–541), Moscow: MSLU.

Reich, A. R. & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *Journal of the Acoustical Society of America, 66,* 1023–1028.

Robertson, B. & Vignaux, G. A. (1995). *Interpreting Evidence*. Chichester: Wiley.

Rose, P. & Duncan, S. (1995). Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics, 2(1),* 1–17.

Rothman, H. B. (1977). A perceptual (aural) and spectrographic identification of talkers with similar sounding voices. *Second international conference, Crime countermeasures-sciences and engineering*. Oxford, England.

Schiller, N. O. & Köster, O. (1996). Evaluation of a foreign language speaker in forensic phonetics: A report. *Forensic Linguistics, 3,* 176–185.

Schötz, S. (2006). *Perception, analysis and synthesis of speaker age* (Published doctoral thesis). Lund: Media-Tryck.

Skarnitzl, R. et al. (2012). Užitečnost formantových kontur při rozpoznávání mluvčích. Unpublished seminar paper, Univerzita Karlova, Praha.

Skarnitzl, R., Vaňková, J., & Bořil, T. (2014). Optimizing formant extraction in Praat and Snack: Comparison of manual and automatic measurements. Presented on *22nd Czech-German Workshop on Speech Communication*. Praha.

Sounddogs.com. (2014). Retrieved from http://www.sounddogs.com

Stevens, K. N. et al. (1968). Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *Journal of the Acoustical Society of America, 44,* 1596–1607.

Syntrillium Software Corporation (2002). Cool Edit Pro (Version 2.00) [Software]. Phoenix.

Thompson, C. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1,* 121–131.

Tiersma, P. & Solan, L. (2002). The linguist on the witness stand: forensic linguistics in American courts. *Language, 78,* 221–239.

Tosi, O. et al. (1972). Experiment on voice identification. *Journal of the Acoustical Society of America, 51*, 2030–2043.

Vaňková, J. (2013). The efficiency of different formant parameters for speaker discrimination. *Phonetica Pragensia XIII*, *AUC Philologica, 1/2014,* 43–54.

Wolf, J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America, 51*, 2044-2056.

Yarmey, A. D. (1991). Voice identification over the telephone. *Journal of Applied Social Psychology, 21:* 1868–1876.

Young, M. A. & Campbell, R. A. (1967). Effects of context on talker identification. *Journal of the Acoustical Society of America, 42*, 1250–1254.

Seznam příloh

Všechny přílohy k této práci se nacházejí na vloženém CD-ROMu. Složka Šumy obsahuje ukázky všech šumů použitých v nahrávkách. Soubor programu Statistica obsahuje všechny vytvořené histogramy rozdělení LTF hodnot. Na disku je také nahrána ukázka z rozhovoru jednoho mluvčího a příslušná část vyextrahovaného vokalického proudu:

složka Šumy

- Bílý šum č. 1. wav
- Bílý šum č. 2.wav
- Hnědý šum č. 1.wav
- Hnědý šum č. 2.wav
- Kavárenský šum č. 1.wav
- Kavárenský šum č. 2.wav
- Kavárenský šum č. 3.wav

LTF – histogramy.stw

Ukázka nahrávky.wav

Ukázka vokalického proudu.wav